Estimation of a Survival Curve with Randomly Censored Data in the Presence of a Covariate

Uwe Jensen, Jürgen Wiedmann

Department of Stochastics, University of Ulm, 89081 Ulm, Germany
(e-mail: jensen@mathematik.uni-ulm.de, wiedmann@mathematik.uni-ulm.de)

Abstract

This paper deals with the estimation of a survival curve in models with random right censoring and dependent censoring mechanism. We consider a specific dependent censorship model in which conditional on a covariate, the survival and censoring times are assumed to be independent. We investigate asymptotic properties of a corrected version of a survival curve estimator introduced by Cheng (1989). In particular we show uniform strong consistency and weak convergence to a Gaussian process. Comparisons of this estimator with the well-known Kaplan-Meier-estimator are included. Finally, some examples illustrate how the estimator performs.

Key words: Nonparametric survival curve estimator, informative random censoring, Kaplan-Meier-estimator, conditional independence model

1 Introduction

In reliability and survival analysis one is often faced with censored lifetime data, i.e. with only partially observable lifetimes. We consider the following well-known model for randomly right censored data.

Let (T, U), $(T_1, U_1), ..., (T_n, U_n)$ be independent pairs of positive random variables where T_j represents the lifetime (failure time) and U_j the censoring time of the j-th object under study. In a model of right random censoring the observations consist of the pairs:

$$(X_j, \delta_j), \quad j = 1, ..., n,$$

where $X_j = \min(T_j, U_j)$ and $\delta_j = I(T_j \leq U_j)$ is an indicator showing whether X_j is the failure ($\delta_j = 1$) or the censoring time ($\delta_j = 0$). Further let F and G denote the distribution functions of the lifetime T and of the censoring variable U, respectively. Here and in the following for any distribution function G we denote by $\bar{G}$ the survival or tail function $\bar{G}(\cdot) = 1 - G(\cdot)$.
In a model with random right censoring the problem of interest is the estimation of the survival function

$$1 - F(t) = \bar{F}(t) = P(T > t).$$

The cumulative hazard function

$$\Lambda(t) = \int_0^t \frac{dF(u)}{\bar{F}(u-)}.$$

uniquely determines the distribution by the relation

$$\bar{F}(t) = \exp\{-\Lambda^c(t)\} \prod_{s \leq t} (1 - \Delta\Lambda(s)) \tag{1}$$

for all t such that $\Lambda(t) < \infty$, where $\Delta\Lambda(s) = \Lambda(s) - \Lambda(s-)$ is the jump height at time s and $\Lambda^c(t) = \Lambda(t) - \sum_{s \leq t} \Delta\Lambda(s)$ is the continuous part of Λ. Therefore, an estimator for $\bar{F}(t)$ can be derived from an estimator for $\Lambda(t)$.

Under the assumption that T and U are independent, the well-known nonparametric survival curve estimator introduced by Kaplan and Meier (1958) has the property of uniform strong consistency for $\bar{F}$ on the support of the distribution of $X = T \wedge U$ (see among others Stute and Wang (1993), Chen and Lo (1997) and the cited literature). This KM-estimator or Product-Limit estimator, is defined by:

$$\hat{\bar{F}}_n^{KM}(t) = \prod_{i: X_{(i)} \leq t} \left(1 - \frac{\delta_{(i)}}{n - i + 1}\right),$$

where $X_{(1)} \leq ... \leq X_{(n)}$ are the order statistics of $X_1, ..., X_n$ and $\delta_{(i)}$ is the censoring indicator corresponding to $X_{(i)}$. However, the assumption of an independent censoring mechanism, often referred to as non-informative censoring, is not always justified. In some cases the censoring variable U or its distribution G carries information about the lifetime (distribution). Such an informative censoring can, for instance, be described by the function $m(x) = P(\delta = 1 | X = x)$. Dikta (1998) assumes that T and U are independent and that m belongs to a parametric class of functions. Then m relates

the distributions G and F to each other so that the observation of a censoring time contributes directly to the estimation of F. If especially m is a constant, then we arrive at the Koziol-Green model in which the survival function $\bar{G}$ is a power of the survival function $\bar{F}$ of the lifetimes, i.e.

$$\bar{G}(t) = \bar{F}(t)^{\beta}, t \geq 0$$

for some fixed, unknown constant $\beta > 0$. This latter model has been reviewed by Csörgö (1988) and Csörgö and Faraway (1998) and generalized by Gather and Pawlitschko (1998). Testing for the Gather-Pawlitschko model was considered by Csörgö (1998). As to be expected estimators under these types of informative censoring outperform the KM-estimator.

In some cases it is not reasonable to assume independence between the censoring variable U and the lifetime T. There may be dependencies due to a covariate Z. For instance, in a competing-risk problem, where some technical system fails due to one or more competing causes, one only observes the time to failure of the system and the corresponding failure mode. If a system with two failure modes A and B fails due to cause A, then the failure time of mode B is randomly censored and vice versa. Since the failure times due to both modes are affected by the same stress and operating environment described by a covariate, it is likely that the failure times are positively correlated. In clinical studies the survival and censoring times can be affected by a set of patient's covariates as age, blood pressure, cholesterol,....

In the model to be considered here we assume that the survival time T and the censoring time U are conditionally independent given a covariate Z. This model has been investigated, among others, by Beran (1981), Dabrowska (1987) and Cheng (1989). The aim of this paper is to present a corrected version of Cheng's estimator of the survival curve and to investigate its asymptotic properties. We make no further assumptions about the specific influence of the covariate. It turns out the even in cases when the KM-estimator is consistent the proposed estimator can perform better than the KM-estimator with respect to the asymptotic variance.

In the following set $\tau = \inf\{t \in \mathbb{R}_+ : \bar{H}(t) = 0\}$ for $H(t) = P(X \leq t)$. It is well known that the KM-estimator may fail to be consistent for $\bar{F}(t)$ in situations in which T and U are dependent. Based on the observations $(X_j, \delta_j), j = 1, .., n$, the following subdistribution functions

$$F_1(t) = P(T_j \leq t, \delta_j = 1), \quad G_0(t) = P(U_j \leq t, \delta_j = 0)$$

can consistently be estimated, and therefore

$$H(t) = P(X_j \leq t) = F_1(t) + G_0(t),$$

too. It is known that $H(t)$ and the subdistributions $F_1(t)$ and $G_0(t)$ do not uniquely determine $\bar{F}(t)$ (see Langberg et al. (1978)). In addition, these authors showed that under the assumption that the discontinuities of $F_1(t)$ and $G_0(t)$ are disjoint on the interval $[0, \tau)$, the KM-estimator in fact consistently estimates the survival function

$$\bar{C}(t) = \exp\left\{ -\int_0^t \frac{dF_1^c(s)}{\bar{H}(s-)} \right\} \times \prod_{s \le t} (1 - \frac{\Delta F_1(s)}{\bar{H}(s-)}) \qquad t \in [0, \tau). \tag{2}$$

Using a result of Williams and Lagakos (1977) or by a direct comparison with (1), it can be shown that the survival function $\bar{C}(t)$ equals $\bar{F}(t)$ under the *constant sum* condition

$$d\Lambda(t) = \frac{dF(t)}{\bar{F}(t-)} = \frac{dF_1(t)}{\bar{H}(t-)},$$

which is fulfilled, in particular, if the survival and censoring times are independent. Models in which this condition fails to hold are called *variable sum* models.

2 The Model

The chosen approach, introduced by Beran (1981) and further developed by Dabrowska (1987) and Cheng (1989), considers in addition to the variables T and U of interest an accompanying covariate Z. The survival time T and the censoring time U are supposed to be conditionally independent given this covariate Z. In the following we present this general set-up following the lines of Cheng (1989).

We assume that the unknown distribution function F of the lifetime T is absolutely continuous with density f. The variables T and U are supposed to be conditionally independent given a covariate Z. For ease of notation we restrict ourselves to a univariate random variable Z; all results with obvious slight modifications can be carried over to the case of a p-dimensional random vector Z. The absolutely continuous distribution function of Z is denoted by R. Thus, the observable data is of the form $(X_j, \delta_j, Z_j), \quad j = 1, \ldots, n$, with

$$X_j = \min(T_j, U_j), \quad \delta_j = I(T_j \le U_j).$$

The aim is to estimate the underlying survival function $\bar{F}$ based on the random sample (X_j, δ_j, Z_j), $j = 1, \ldots, n$.

The starting point here is the conditional cumulative hazard function at time point t given $Z = z$:

$$\Lambda(t|z) = \int_0^t \frac{dF(u|z)}{\bar{F}(u - |z)},$$

where $F(u|z) = P(T \leq u|Z = z)$. Under the conditional independence assumption $\Lambda(t|z)$ and

$$\int_0^t \bar{G}(u-|z)dF(u|z) = P(X \leq t, \delta = 1|Z = z), \quad t \geq 0$$

are identifiable, where $G(u|z) = P(U \leq u|Z = z)$. With $r(\cdot)$ denoting the density function of the distribution of Z, $\Lambda(t|z)$ can be expanded as follows:

$$\Lambda(t|z) = \int_0^t \frac{r(z)\bar{G}(u-|z)dF(u|z)}{r(z)\bar{G}(u-|z)\bar{F}(u-|z)} = \int_0^t \frac{dA(u;z)}{B(u;z)}.$$

Here $A(u;z) = r(z)P(X \leq u, \delta = 1|Z = z)$ and $B(u;z) = r(z)P(X \geq u|Z = z)$. The self suggesting idea now is to estimate A and B by their empirical counterparts A_n and B_n using appropriate kernel estimators:

$$\begin{aligned} A_n(u;z) &= \frac{1}{n}\sum_{i=1}^n I(X_i \leq u, \delta_i = 1)K_b(z, Z_i) \\ B_n(u;z) &= \frac{1}{n}\sum_{i=1}^n I(X_i \geq u)K_b(z, Z_i), \end{aligned}$$

where $K_b(u,v) = \frac{1}{b}K(\frac{u-v}{b})$ with a kernel function K and bandwidth $b = b_n$ tending to 0 as $n \to \infty$. As a result we get the following estimators for the conditional cumulative hazard functions

$$\Lambda_n(t|z) = \int_0^t \frac{dA_n(u;z)}{B_n(u;z)} \tag{3}$$

and for the conditional survival functions $\bar{F}(t|z)$

$$\bar{F}_n(t|z) = \exp\left\{-\Lambda_n(t|z)\right\}.$$

To show uniform strong consistency of $\Lambda_n(t|z)$ and $\bar{F}_n(t|z)$ the following condition is needed (see Cheng (1989), Dabrowska (1989) and Rosenblatt (1971)):

Condition 1

1.1 The kernel function K is bounded and Lipschitz continuous of order 1 with respect to the Euclidean distance on $\mathbb{R}$.

1.2 $\int K(z)dz = 1$, $\int zK(z)dz = 0$ and $\int z^2 |K(z)|\, dz < \infty$.

1.3 The bandwidth sequence $b \equiv b_n$ fulfills: $b \to 0$ and $\frac{\log n}{nb} \to 0$ $(n \to \infty)$.

1.4 The partial derivatives, with respect to t, of $F(t|z)$ and $G(t|z)$ exist and are continuous in t for each z.

1.5 The functions $r(z)$, $F(t|z)$ and $G(t|z)$ have bounded continuous first and second partial derivatives with respect to z.

1.6 For any closed interval $[a,d] \subset \mathbb{R}_+$, there exist constants $\rho, \delta(\varepsilon) > 0$ such that $P(X > \rho | Z = z) \geq \delta(\varepsilon), \quad \forall z \in [a,d]$ with $r(z) \geq \varepsilon$, $\varepsilon > 0$ arbitrarily small.

Note that the last part of this condition implies that

$$P(X > \rho) \geq \varepsilon \int_{z:r(z)\geq\varepsilon} P(X > \rho | Z = z) dz > 0.$$

Cheng (1989) showed that Condition 1 ensures asymptotic unbiasedness of A_n and B_n and uniform strong consistency of $\Lambda_n(t|z)$ and $\bar{F}_n(t|z)$ in the following sense

$$\sup_{0\leq t\leq\tau^\star} \sup_{z\in[a,d]} |\Lambda_n(t|z) - \Lambda(t|z)| = O(b^2) + O\left(\left(\frac{\log n}{nb}\right)^{1/2}\right) \quad a.s. \tag{4}$$

$$\sup_{0\leq t\leq\tau^\star} \sup_{z\in[a,d]} |\bar{F}_n(t|z) - \bar{F}(t|z)| = O(b^2) + O\left(\left(\frac{\log n}{nb}\right)^{1/2}\right) \quad a.s. \tag{5}$$

over a rectangle of the form $[0, \tau^\star] \times [a,d]$, where $a, d \in \mathbb{R}$ with $a \leq d$ and $\tau^\star \in (0, \tau)$.

Remark 1 *Under the additional assumptions that* $\sum b_n^\xi < \infty$ *for some* $\xi > 0$ *and* $nb_n^5 \to 0$, *Dabrowska* (1989) *showed strong uniform consistency at a rate* $O(\sqrt{\frac{\log b_n^{-1}}{nb_n}})$. *The choice of* $b_n = n^{-\alpha}$ *demands that for strong uniform consistency* $\frac{1}{5} < \alpha < 1$. *For all such permissible values of* α *this results in a better rate of convergence than the rate stated above.*

2.1 Asymptotic Behaviour of an Estimator of the Survival Function

Denoting R_n the empirical distribution function based on the sample $(Z_1, ..., Z_n)$ of covariates, Cheng (1989) used

$$\widetilde{\Lambda}_n(t) = \int \Lambda_n(t|z) dR_n(z)$$

as an estimator for the cumulative hazard function $\Lambda(t)$ and consequently $\widetilde{\bar{F}}_n(t) = \exp\{-\widetilde{\Lambda}_n(t)\}$ as an estimate of $\bar{F}(t)$. We show that $\widetilde{\Lambda}_n(t)$ in fact estimates an upper bound of $\Lambda(t)$ and therefore, $\widetilde{\Lambda}_n(t)$ and $\widetilde{\bar{F}}_n(t)$ are not consistent in general. In this section we introduce a modified (corrected) version of the estimator proposed by Cheng and investigate its asymptotic properties.
Define the convex function $g(x) = -\ln(1-x)$, $0 \leq x < 1$. Then we obtain:

$$\begin{aligned}
\Lambda(t) &= g(F(t)) = g(EI(T \leq t)) = g(EE[I(T \leq t)|Z]) \\
&\leq Eg(E[I(T \leq t)|Z]) \\
&= \int \Lambda(t|z) dR(z).
\end{aligned}$$

Recall that we assumed that F is absolutely continuous. This inequality shows that Cheng's considerations deal with upper bound estimators for the true unconditional cumulative hazard function. As a direct consequence the proposed estimator for the survival curve $\widetilde{\bar{F}}_n(t) = \exp\{-\widetilde{\Lambda}_n(t)\}$ is an estimator of a lower bound of $\bar{F}(t)$. Cheng (1989) showed, as the main result of his paper, weak convergence of the process $\sqrt{n}(\widetilde{\Lambda}_n(t) - \int \Lambda(t|z) dR(z))$ to a centered Gaussian process ($t \in [0, \tau^\star]$).
Starting with $\bar{F}_n(t|z)$, as proposed by Cheng, we define the modified estimator

$$\widehat{\bar{F}}_n(t) = \int \exp\left(-\Lambda_n(t|z)\right) dR_n(z) = \int \bar{F}_n(t|z) dR_n(z),$$

which we will call CIM-estimator (Conditional Independence Model) in the following. We will show that $\widehat{\bar{F}}_n(t)$ is a consistent estimator for the unconditional survival function $\bar{F}(t)$.

Theorem 1 *Under Condition* 1 $\widehat{\bar{F}}_n(t)$ *is a uniformly strongly consistent estimator for* $\bar{F}(t)$ *on* $[0, \tau)$, *i.e.* $\sup_{0 \leq t < \tau} |\widehat{\bar{F}}_n(t) - \bar{F}(t)| \to 0$ *as* $n \to \infty$, *a.s.*

Proof. We will use the following decomposition of $\widehat{\bar{F}}_n(t) - \bar{F}(t)$:

$$\begin{aligned}
\widehat{\bar{F}}_n(t) - \bar{F}(t) &= \int \bar{F}_n(t|z) dR_n(z) - \int \bar{F}(t|z) dR(z) \\
&= \int \left(\bar{F}_n(t|z) - \bar{F}(t|z)\right) dR_n(z) + \int \bar{F}(t|z) d\left(R_n(z) - R(z)\right) \\
&= C_n + D_n
\end{aligned}$$

If the distribution of Z has a bounded support covered by $[a,d]$, then (5) ensures that

$$|C_n| \leq \sup_{z\in[a,d]} |\bar{F}_n(t|z) - \bar{F}(t|z)| = O(b^2) + O(\sqrt{\frac{\log n}{nb}}). \tag{6}$$

Otherwise, we split C_n into two integrals $\int_a^d + \int_d^\infty$. For the second integral we can choose d large enough such that for a given $\varepsilon > 0$ the term $|C_n^{(2)}|$, given by

$$|C_n^{(2)}| = \left| \int_d^\infty \left(\bar{F}_n(t|z) - \bar{F}(t|z)\right) dR_n(z) \right| \leq 2\frac{1}{n}\sum_{i=1}^n I(Z_i > d) \to 2\bar{R}(d),$$

can be bounded by ε a.s. for all sufficiently large n. For the first term we may again apply (6) yielding $|C_n| \to 0$ a.s.
D_n can be rewritten as

$$\int \bar{F}(t|z) d\left(R_n(z) - R(z)\right) = \frac{1}{n}\sum_{i=1}^n \bar{F}(t|Z_i) - \bar{F}(t).$$

The SLLN ensures $|D_n| \to 0$ as $n \to \infty$, a.s.
Since $\hat{\bar{F}}_n(t)$ and $\bar{F}(t)$ are monotone functions, pointwise convergence can be carried over to uniform convergence on $[0,\tau^\star] \quad \forall \tau^\star < \tau$ (cf. Shorack and Wellner (1986), p. 96). Since $\bar{F}$ is bounded, this extends to uniform convergence on $[0,\tau)$. ■

To extend this result the common way is to represent the difference $D_n(t) = \hat{\bar{F}}_n(t) - \bar{F}(t)$ as a sum of i.i.d. random variables and a remainder term similar to Lo and Singh (1986), Csörgö (1996) or Gather and Pawlitschko (1998).

Proposition 2 *Let the distribution of Z be concentrated on a finite interval $[a,d]$ and define $M(t,z) = \bar{F}(t|z) - \bar{F}(t)$. Then under Condition* 1

$$D_n(t) = \frac{1}{n}\sum_{i=1}^n M(t,Z_i) + V_n(t), \quad 0 \leq t < \tau,$$

where for any $\tau^\star \in (0,\tau)$ with probability one

$$\sup_{0\leq t\leq\tau^\star} |V_n(t)| = O(b^2) + O(\sqrt{\frac{\log n}{nb}}).$$

Proof. Consider the following decomposition of $D_n(t)$:

$$D_n(t) = D_n^{(1)}(t) + D_n^{(2)}(t) + D_n^{(3)}(t) \tag{7}$$

with

$$\begin{aligned} D_n^{(1)}(t) &= \int (\bar{F}_n(t|z) - \bar{F}(t|z)) dR(z), \\ D_n^{(2)}(t) &= \int \bar{F}(t|z) d\left[R_n(z) - R(z)\right], \\ D_n^{(3)}(t) &= \int (\bar{F}_n(t|z) - \bar{F}(t|z)) d\left[R_n(z) - R(z)\right]. \end{aligned}$$

1. For the first term $D_n^{(1)}(t)$ we get with $\Delta_n(t|z) = \Lambda_n(t|z) - \Lambda(t|z)$:

$$\begin{aligned} \bar{F}_n(t|z) - \bar{F}(t|z) &= \bar{F}(t|z)(\exp\{-\Delta_n(t|z)\} - 1) \\ &= \bar{F}(t|z)\left(-\Delta_n(t|z) + o(\Delta_n(t|z))\right) \\ &= \bar{F}(t|z)\Delta_n(t|z)\left(o(1) - 1\right). \end{aligned}$$

   Using (5) it follows for $n \to \infty$:

$$\sup_{0 \le t \le \tau^\star} \sup_{z \in [a,d]} |\bar{F}(t|z)\Delta_n(t|z)| = O(b^2) + O(\sqrt{\frac{\log n}{nb}}).$$

   This yields

$$\sup_{0 \le t \le \tau^\star} |D_n^{(1)}(t)| = O(b^2) + O(\sqrt{\frac{\log n}{nb}}).$$

2. The second term $D_n^{(2)}(t)$ is:

$$D_n^{(2)}(t) = \int \bar{F}(t|z) dR_n(z) - \bar{F}(t) = \frac{1}{n}\sum_{i=1}^{n} M(t, Z_i).$$

3. Using uniform strong convergence of $\bar{F}_n(t|z)$ to $\bar{F}(t|z)$, the dominated convergence theorem and noting that $R_n(z)$ and $R(z)$ are bounded monotone functions, it follows that $D_n^{(3)}(t)$ converges to zero uniformly with probability one at the rate $O(b^2) + O(\sqrt{\frac{\log n}{nb}})$.

Combining the results of these three cases proves the assertion. ■

Investigating the first part $D_n^{(1)}(t)$ of the decomposition of $D_n(t)$ in the proof of Proposition 2 carefully allows to separate a further additive term yielding

$$D_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(L(t, Z_i, X_i, \delta_i) + M(t, Z_i)\right) + \tilde{V}_n(t), \quad 0 \le t < \tau,$$

where for $t, z, x \ge 0, \delta \in \{0, 1\}$

$$L(t, z, x, \delta) = \bar{F}(t|z)\left(\int_0^t \frac{I(x > u)dF(u|z)}{\bar{G}(u|z)\bar{F}^2(u|z)} - \frac{I(x \le t, \delta = 1)}{\bar{F}(x|z)\bar{G}\,(x|z)}\right). \tag{8}$$

We omit a detailed derivation of this result here (compare the proof of Theorem 3) since we do not gain a better rate of convergence of the remainder $\tilde{V}_n(t)$ compared to that in Proposition 2. Unfortunately this rate of convergence is not good enough to investigate the rate of strong uniform consistency of the estimator $\hat{\bar{F}}_n(t)$ or to establish weak convergence. So this has to be carried out via alternative methods, for which we make use of the following condition.

Condition 2

2.1 $$E\left(\bar{F}^2(t|Z)\int_0^t \frac{dF(u|Z)}{\bar{G}(u|Z)\bar{F}^2(u|Z)}\right) < \infty \quad \forall t \in [0, \tau).$$

2.2 $$\sqrt{n}\cdot b^2 \longrightarrow 0, \frac{\log n}{\sqrt{n}b} \longrightarrow 0 \quad (n \to \infty).$$

To prove a weak convergence result for the CIM-estimator $\hat{\bar{F}}_n(t)$, we have to show weak convergence of the process:

$$W_n(t) = \sqrt{n}\left(\int \bar{F}_n(t|z)dR_n(z) - \bar{F}(t)\right)$$

on the space $\mathcal{D}[0, \tau^\star]$ of right continuous functions on $[0, \tau^\star]$ with finite left-hand limits (cadlag functions) equipped with the Skorohod topology. We note that the recent best weak-convergence results of Csörgö (1996) for the KM-estimator allow data intervals increasing to $[0, \tau]$. Our technique restrict us to a fixed interval $[0, \tau^\star]$ and thus leaves the question of corresponding extensions open.

Theorem 3 *Under Conditions* 1 *and* 2 *the process* $(W_n(t)), t \in \mathbb{R}_+$, *converges weakly on* $\mathcal{D}[0, \tau^\star]$ *for* $0 \leq \tau^\star < \tau$ *to a mean zero Gaussian process* $W = (W(t)), t \in \mathbb{R}_+$, *with covariance function* $(0 \leq s \leq t \leq \tau^\star)$*:*

$$\text{Cov}\,(W(s), W(t)) = E\left(\bar{F}(s|Z)\bar{F}(t|Z)\left[\int_0^s \frac{dF(u|Z)}{\bar{G}(u|Z)\bar{F}^2(u|Z)} + 1\right]\right) - \bar{F}(s)\bar{F}(t). \quad (9)$$

Proof. The proof uses similar methods as presented in Cheng (1989) and will be given in 3 steps.

1. We use the same decomposition as in the proof of Proposition 2:
   $W_n = W_n^{(1)} + W_n^{(2)} + W_n^{(3)}$, where $W_n^{(i)} = \sqrt{n} D_n^{(i)}, i = 1, 2, 3$, and

$$\begin{aligned} D_n^{(1)}(t) &= \int (\bar{F}_n(t|z) - \bar{F}(t|z)) dR(z), \\ D_n^{(2)}(t) &= \int \bar{F}(t|z) d\left[R_n(z) - R(z)\right], \\ D_n^{(3)}(t) &= \int (\bar{F}_n(t|z) - \bar{F}(t|z)) d\left[R_n(z) - R(z)\right]. \end{aligned}$$

   For the term $W_n^{(1)}(t)$ we recall from the proof of Proposition 2 that

$$\bar{F}_n(t|z) - \bar{F}(t|z) = \bar{F}(t|z)\Delta_n(t|z)\left(o(1) - 1\right)$$

   where $\Delta_n(t|z) = \Lambda_n(t|z) - \Lambda(t|z)$. So the asymptotic distribution properties of $W_n^{(1)}$ coincide with those of

$$-n^{1/2} \int \bar{F}(t|z)\Delta_n(t|z) dR(z).$$

   Here we can directly use (3.6) in Cheng (1989) to see that uniformly on $[0, \tau^\star]$,

$$\begin{aligned} W_n^{(1)}(t) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \bar{F}(t|Z_i) \left\{ \int_0^t \frac{I\left(X_i > u\right) dA(u; Z_i)}{B^2\left(u; Z_i\right)} - \frac{I\left(X_i \leq t, \delta_i = 1\right)}{B\left(X_i; Z_i\right)} \right\} r(Z_i) \\ &\quad + O(n^{1/2} b^2) \end{aligned}$$

   in probability for $n \to \infty$.

$W_n^{(2)}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\bar{F}(t|Z_i) - \bar{F}(t))$ is a sum of i.i.d. random variables with zero mean $E(\bar{F}(t|Z) - \bar{F}(t)) = 0$. Therefore the CLT applies to yield joint asymptotic normality for all finite dimensional marginals.

$W_n^{(3)}(t)$ converges to zero in probability for each fixed $t \in [0, \tau^\star]$: Tightness of the probability measures associated with the process $\rho_n(z) = \sqrt{n}\,[R_n(z) - R(z)]$ is well known (Billingsley (1968), Theorem 16.4) and the uniform strong convergence of $\bar{F}_n(t|z)$ to $\bar{F}(t|z)$ on finite intervals $[a, d]$ follows from (5). This yields, following Cheng (1989) that $W_n^{(3)}(t) \xrightarrow{P} 0$ as $n \to \infty$.

2. The conclusion from these considerations concerning $W_n^{(i)}(t)$ is that the asymptotic distribution of $W_n(t)$ is that of

$$W_n^*(t) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} (L_i(t) + M_i(t)),$$

where $L_i(t) = L(t, Z_i, X_i, \delta_i)$ and $M_i(t) = M(t, Z_i)$ are defined as before in Proposition 2 and (8). To determine the covariance structure of $W(t)$ we notice after some straightforward calculations that for $0 \le s \le t \le \tau^\star$:

$EL_i(t) = 0$, $EM_i(t) = E\left(\bar{F}(t|Z_i) - \bar{F}(t)\right) = 0$, $EL_i(s)M_i(t) = EM_i(s)L_i(t) = 0$ and

$$EL_i(s)L_i(t) = E\left(\bar{F}(s|Z_i)\bar{F}(t|Z_i) \int_0^s \frac{dF(u|Z_i)}{\bar{G}(u|Z_i)\bar{F}^2(u|Z_i)}\right),$$

where the latter result can either be achieved by direct computation or by using Proposition 2.3 of Dabrowska (1987). Furthermore, we have $EM_i(s)M_i(t) = E\bar{F}(s|Z_i)\bar{F}(t|Z_i) - \bar{F}(s)\bar{F}(t)$. By means of the CLT this shows that the finite dimensional marginal distributions of $W_n^*(t)$ are asymptotically normal with zero mean and covariance function specified by (9).

3. It remains to show tightness for the sequence of distributions associated with the random functions $W_n^*(t)$. For this we will use condition (15.21) of Billingsley (1968). We have to establish that for $0 \le t_1 \le t \le t_2 \le \tau^\star$, $n \ge 1$, and a non-decreasing, continuous function Q, the following inequality holds:

$$E\left(|W_n(t) - W_n(t_1)|^2 \, |W_n(t_2) - W_n(t)|^2\right) \le [Q(t_2) - Q(t_1)]^2 .$$

This amounts to showing that such an upper bound applies to

$$A_L = \frac{1}{n^2} E\left(\left(\sum_{i=1}^{n} L_{1i}\right)^2 \left(\sum_{i=1}^{n} L_{2i}\right)^2\right)$$

and

$$A_M = \frac{1}{n^2} E\left(\left(\sum_{i=1}^{n} M_{1i}\right)^2 \left(\sum_{i=1}^{n} M_{2i}\right)^2\right)$$

with

$$\begin{aligned} L_{1i} &= L_i(t) - L_i(t_1), \quad M_{1i} = M_i(t) - M_i(t_1) \\ L_{2i} &= L_i(t_2) - L_i(t), \quad M_{2i} = M_i(t_2) - M_i(t). \end{aligned}$$

For the term A_L this can directly be deduced from (3.9) in Cheng (1989). For the term A_M we can write $A_M = E\left(n(\bar{M}(t) - \bar{M}(t_1))^2(\bar{M}(t_2) - \bar{M}(t))^2\right)$ with $\bar{M}(t) = \frac{1}{n}\sum_{i=1}^{n} M_i(t)$. Following the argumentation of Lo and Singh (1985) and noting that the random variables $M_i(t)$ are bounded uniformly in $[0, \tau^\star]$ with $EM_i(t) = 0$ and known covariance structure (see above) it follows that for all $s, t \in [0, \tau^\star], \mathrm{Var}(M_1(s) - M_1(t)) \leq |Q_M(s) - Q_M(t)|$ for a continuous non-decreasing function Q_M. This leads to

$$\begin{aligned} A_M &\leq 3E(M_1(t) - M_1(t_1))^2 E(M_1(t_2) - M_1(t))^2 \\ &\leq 3\left(Q_M(t_2) - Q_M(t_1)\right)^2, \end{aligned}$$

which completes the proof. ■

2.2 Confidence Bounds for the Survival Curve

With the results from Theorem 3 we are now able to derive pointwise confidence intervals for $\bar{F}(t)$. Using a consistent estimator $\widehat{v}_n(t)$ for the asymptotic variance

$$v(t) = E\left[\bar{F}^2(t|Z)\left(\int_0^t \frac{dF(u|Z)}{\bar{G}(u|Z)\bar{F}^2(u|Z)}\right)\right] + \mathrm{Var}(\bar{F}(t|Z))$$

of the process W we obtain pointwise two-sided asymptotic confidence intervals for $\bar{F}(t)$ to the level $(1-\alpha), 0 < \alpha < 1$:

$$\widehat{\bar{F}}_n(t) \pm \sqrt{\frac{\widehat{v}_n(t)}{n}} z_{1-\alpha/2},$$

with $z_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the $N(0,1)$-distribution. Using the plug-in-technique, a consistent estimator $\widehat{v}_n(t)$ of $v(t)$ can be found to be

$$\widehat{v}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\bar{F}_n^2(t|Z_i) r_n^2(Z_i) I(X_i \le t, \delta_i = 1)}{B_n^2(X_i; Z_i)} + \frac{1}{n}\sum_{i=1}^{n} \left(\bar{F}_n(t|Z_i) - \overline{\bar{F}_n(t|Z)}\right)^2,$$

where $r_n(z) = \frac{1}{n}\sum_{i=1}^{n} K_b(z, Z_i)$ is the usual kernel density estimator for $r(z)$ and $\overline{\bar{F}_n(t|Z)} = \frac{1}{n}\sum_{j=1}^{n} \bar{F}_n(t|Z_j)$. If the covariate Z is degenerate, i.e. a constant, then $v(t)$ reduces to $\bar{F}^2(t)\int_0^t \frac{dF(u)}{\bar{G}(u)\bar{F}^2(u)}$, the well known asymptotic variance of the Kaplan-Meier estimator (cf. Breslow and Crowley (1974)).

2.3 Estimation with Discrete Covariates

A covariate with an absolutely continuous distribution needs density estimation with kernel smoothing. Less expenditure is necessary if the covariate takes only finitely many values, say $1, ..., m$, as often applies in practice, e.g. in cases where covariates such as sex, type of treatment, ... are involved. For such discrete covariates the classical KM-estimator within each class $\{Z = k\}, k = 1, ..., m$, can be used to estimate the underlying law of F, which can be expressed as

$$\bar{F}(t) = \sum_{k=1}^{m} p_k \bar{F}(t|Z = k)$$

with weights $p_k = P(Z = k)$. The obvious idea is to use KM-estimates $\widehat{\bar{F}}_n^{KM}(t|Z = k)$ for $\bar{F}(t|Z = k)$ in the stratified sample and to estimate the weights p_k by

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} I(Z_i = k).$$

The resulting estimate

$$\widehat{\bar{F}}_d^{KM}(t) = \sum_{k=1}^{m} \hat{p}_k \widehat{\bar{F}}_n^{KM}(t|Z = k) \tag{10}$$

is uniformly strongly consistent as the following proposition shows.

Proposition 4 *In the discrete covariate model the estimator* $\widehat{\bar{F}}_d^{KM}(t)$ *defined in* (10) *is uniformly strongly consistent for* $\bar{F}(t)$ *on* $0 \le t \le \tau^\star, \tau^\star \in (0, \tau)$, *where* $\widehat{\bar{F}}_n^{KM}(t|Z=k)$ *is the KM-estimate of the subsample* $\{(X_{i_j}, \delta_{i_j}, Z_{i_j}) : Z_{i_j} = k,\ j = 1, \ldots, n_k\}, \sum_{k=1}^m n_k = n, m \le n$, *provided that* n_k *tends to* ∞ *as* $n \to \infty$.

Proof. From the SLLN we have that $\hat{p}_k$ converges to p_k with probability 1. Also we have the well known result of uniform strong consistency of $\widehat{\bar{F}}_n^{KM}(t|Z=k)$ for $t \in [0, \tau^\star]$. This implies pointwise strong consistency of $\widehat{\bar{F}}_d^{KM}(t)$ for $0 \le t \le \tau^\star$, i.e.

$$\widehat{\bar{F}}_d^{KM}(t) = \sum_{k=1}^m \hat{p}_k \widehat{\bar{F}}_n^{KM}(t|Z=z_k) \to \bar{F}(t)$$

with probability 1. Since $\bar{F}$ and $\widehat{\bar{F}}_d^{KM}$ are decreasing functions this implies uniform strong consistency of $\widehat{\bar{F}}_d^{KM}(t)$ on $[0, \tau^\star]$. ■

As an example we consider a model with a covariate which takes only two values; for brevity we omit the details of the model here. Figure 1 shows the results of a simulation. In this example the constant-sum condition fails to hold. Therefore, the crude KM-estimator is consistent for $\bar{C}(t)$ as defined in (2), which deviates from $\bar{F}(t)$. In addition Figure 1 shows the two KM-estimates of the stratified sample $\widehat{\bar{F}}_n^{KM}(t|Z=k), k=1,2$ and the linear combination $\widehat{\bar{F}}_d^{KM}(t)$, which fits $\bar{F}$ well.

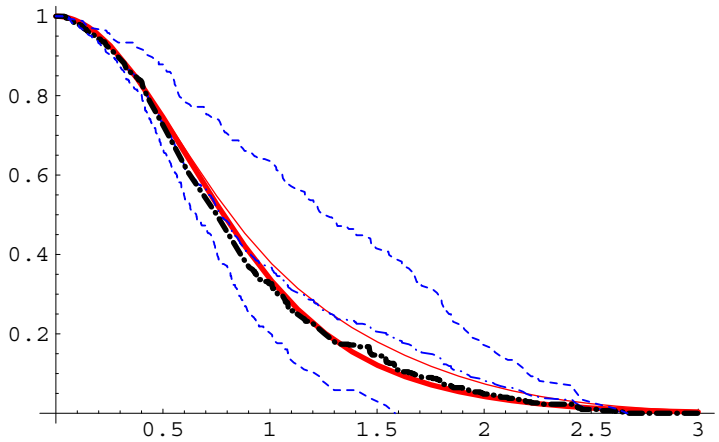

Figure 1: Estimates $\widehat{\bar{F}}_n^{KM}(t|Z=k)$, $k=1,2$ - - -, $\widehat{\bar{F}}_n^{KM}(t)$ –·– and $\widehat{\bar{F}}_d^{KM}(t)$ –·– with functions $\bar{C}(t)$ —— and $\bar{F}(t)$ ——.

Finally, we remark that the asymptotic behaviour of the estimator $\widehat{\bar{F}}_d^{KM}(t)$ can be investigated in more detail. It is possible to represent the difference $\widehat{\bar{F}}_d^{KM}(t) - \bar{F}(t)$ as an average of i.i.d. variables with an additive remainder of order $O(\frac{\log n}{n})$. This implies that the rate of convergence in Proposition 4 is $O\left(\sqrt{\frac{\log\log n}{n}}\right)$. In addition weak convergence to a mean zero Gaussian process can be shown. For details we refer to [10].

3 Examples and an Application

3.1 A Variable-Sum Model

For constant-sum models the KM-estimator $\widehat{\bar{F}}_n^{KM}(t)$ provides a consistent nonparametric estimate for $\bar{F}(t)$. In this section we will introduce a conditional Koziol-Green-type model for which the constant-sum condition is not fulfilled. The aim of this section is to illustrate that under this condition the introduced estimator is still consistent whereas the crude KM-estimator deviates significantly from $\bar{F}(t)$. In this model we define the conditional distributions to be of a Weibull-type with common shape parameter $\theta > 0$ and scale parameters $\alpha, \beta > 0$. Z is assumed to be exponentially distributed with parameter $\lambda > 0$:

$$\begin{aligned} F(t|z) &= 1 - \exp\left(-\alpha z t^{\theta}\right) \\ G(t|z) &= 1 - \exp\left(-\beta z t^{\theta}\right) \\ R(z) &= 1 - \exp\left(-\lambda z\right) \end{aligned}$$

Here the covariate Z has an multiplicative effect on the failure rates of T and U, respectively. The lifetime T and censoring time U are conditional independent given Z. Condition 2.1 is met if $\alpha \geq \beta$. First we show that the constant-sum condition fails to hold in this setting. The unconditional failure rates can be calculated as follows:

$$\begin{aligned} \lambda(t) = \frac{f(t)}{\bar{F}(t)} &= \frac{\alpha\theta t^{\theta-1}}{\alpha t^{\theta} + \lambda} \\ \lambda^{\#}(t) = \frac{f_1(t)}{\bar{H}(t)} &= \frac{\alpha\theta t^{\theta-1}}{(\alpha+\beta)\, t^{\theta} + \lambda} \end{aligned}$$

This implies that $\lambda(t) > \lambda^{\#}(t) \quad \forall t \in (0, \tau^{\star}], \quad \alpha, \beta, \theta > 0$, which shows that the constant-sum condition is not met. As a result the KM-estimator is no longer consistent

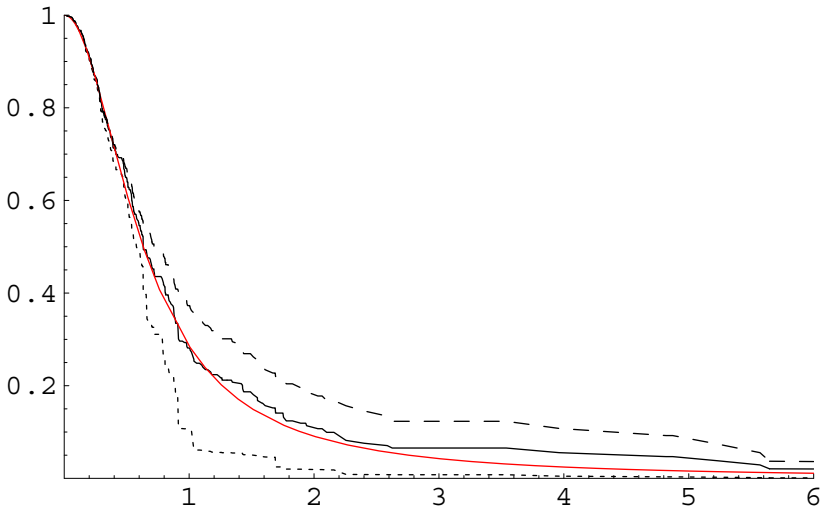

Figure 2: CIM-estimate $\widehat{\bar{F}}_n(t)$ ——, K-M-estimate $\widehat{\bar{F}}_n^{KM}(t)$ - - -, Cheng estimate $\widetilde{\bar{F}}_n(t)$ ···· and true survival curve $\bar{F}(t)$.

for $\bar{F}(t)$. A useful way to examine whether the KM-estimator provides an estimate of an upper or lower bound of $\bar{F}(t)$ is to check the right-tail-increasing (RTI) condition of Nair (1993). If this condition is fulfilled $\bar{C}(t)$ as defined in (2) is an upper bound of $\bar{F}(t)$. This idea seems natural here, because $\lambda^{\#}(t)$, the hazard rate related to the KM-estimate, is strictly smaller than $\lambda(t)$ and so the KM-estimate should estimate an upper bound of $\bar{F}$. The RTI-condition demands that $P(U > u|T > t)$ is non-decreasing in t for every fixed u, which is obviously fulfilled since

$$P(U > u|T > t) = \frac{\alpha t^{\theta} + \lambda}{\alpha t^{\theta} + \beta u^{\theta} + \lambda}.$$

The difference $\lambda(t) - \lambda^{\#}(t)$ is increasing in β for $\alpha, t, \theta, \lambda$ fixed. Since the censoring proportion $P(\delta = 0) = 1 - \lim_{t\to\infty} F_1(t) = \frac{\beta}{\alpha+\beta}$ is solely determined by α and β, the distortion therefore gets larger with higher censoring proportion. Figures 2 and 3 show both estimates and the residuals for $n = 400$, $\alpha = 2.5$, $\beta = 2$, $\lambda = 1$ and $\theta = 2$. The induced theoretical censoring proportion is 44.44%, whereas the sample censoring proportion is 43.25%. We see that the KM and Cheng estimators indeed estimate upper and lower bounds of $\bar{F}$ respectively, whereas the CIM-estimator better fits the true underlying unconditional survival function. This becomes more evident looking at the deviations shown in Figure 3.
Finally we focus on the asymptotic variance $v(t)$ which tends to zero ($t \to \infty$) for all

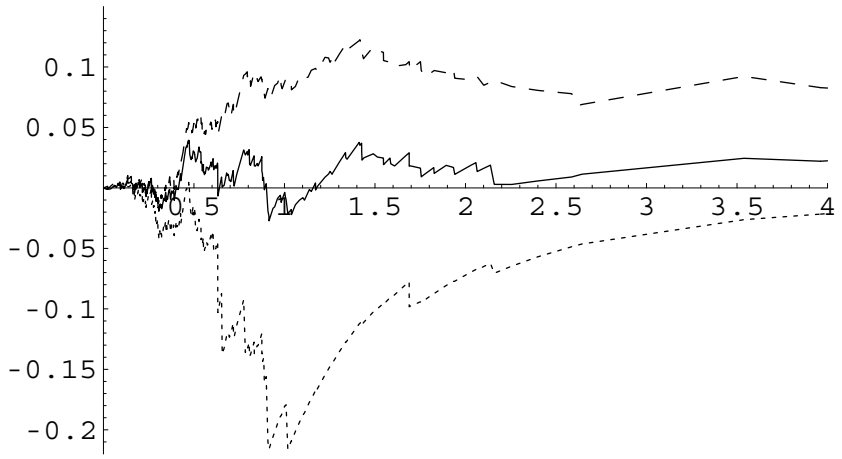

Figure 3: Deviations of CIM-estimate $\widehat{\widetilde{F}}_n(t)$ ——, K-M-estimate $\widehat{\widetilde{F}}_n^{KM}(t)$ - - - and Cheng estimate $\widetilde{\widetilde{F}}_n(t)$ ···· .

$\alpha > \beta$. We note that this condition is equivalent to $P(\delta = 0) < \frac{1}{2}$. Figure 4 illustrates this for the previous parameter constellation.

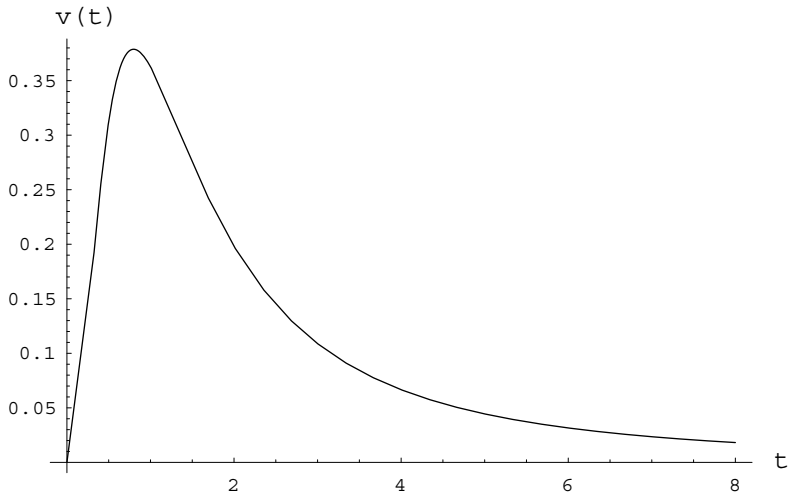

Figure 4: Asymptotic variance $v(t)$ with parameters $\alpha = 2.5, \beta = 2, \lambda = 1, \theta = 2$.

3.2 A Constant-Sum Model

Now we will focus on the asymptotic variance of the process W and compare it with the asymptotic variance of the KM-estimator. Since the model introduced in the previous section seems not appropriate for this purpose (the two estimators estimate different functions) we introduce a model where the constant-sum condition holds true. Let

$$\begin{aligned} F(t|z) &= 1 - \exp\left(-\alpha z t^{\theta}\right) \\ G(t|z) &= 1 - \exp\left(-\beta t^{\theta}\right) \\ R(z) &= 1 - \exp\left(-\lambda z\right) \end{aligned}$$

with $\alpha, \beta, \lambda, \theta > 0$. Here the influence of the covariate is multiplicative only on the hazard rate of the lifetime distribution F. Censoring is not influenced by Z. This approach seems quite natural for many problems in survival and reliability analysis, where some external effect has influence only on the lifetime of the object under study. As before conditional independence of T and U given Z is assumed what in this specific model is equivalent with the assumption that T and U are independent. So the constant-sum condition is fulfilled and the failure rates $\lambda(t)$ and $\lambda^{\#}(t)$ both equal $\frac{\alpha\theta t^{\theta-1}}{\alpha t^{\theta}+\lambda}$. As will turn out, even in this case of independence, it pays to include the covariate in the model with respect to the asymptotic variance (see Proposition 5 below).

From (9) the asymptotic variance of the process W is given by

$$v(t) = E\left(\bar{F}^2(t|Z)\int_0^t \frac{dF(u|Z)}{\bar{G}(u|Z)\bar{F}^2(u|Z)}\right) + \mathrm{Var}\left(\bar{F}(t|Z)\right),$$

whereas we have

$$v^{KM}(t) = \bar{F}^2(t)\int_0^t \frac{dF(u)}{\bar{G}(u)\bar{F}^2(u)}$$

for the KM-estimator (Breslow and Crowley (1974)). For this specific model we get:

$$\begin{aligned} v^{KM}(t) &= \left(\frac{\lambda}{\alpha t^{\theta}+\lambda}\right)^2 \int_0^t \frac{\alpha\lambda\theta u^{\theta-1}}{(\alpha u^{\theta}+\lambda)^2}\frac{1}{e^{-\beta u^{\theta}}(\frac{\lambda}{\alpha u^{\theta}+\lambda})^2}du \\ &= \frac{\lambda\alpha(e^{\beta t^{\theta}}-1)}{\beta(\alpha t^{\theta}+\lambda)^2} \end{aligned}$$

and

$$\mathrm{Var}(\bar{F}(t|Z)) = \frac{\lambda\alpha^2 t^{2\theta}}{(\lambda+\alpha t^{\theta})^2(\lambda+2\alpha t^{\theta})}$$

$$v(t) - \mathrm{Var}\left(\bar{F}(t|Z)\right) \;=\; \lambda\left(\frac{e^{\beta t^\theta}}{\lambda+\alpha t^\theta} - \frac{1}{\lambda+2\alpha t^\theta} - \frac{\beta e^{(2\beta t^\theta+\frac{\beta\lambda}{\alpha})}\Gamma(t)}{\alpha}\right).$$

Here we set

$$\Gamma[t] = \Gamma[0, k(t)+d, 2k(t)+d], \tag{11}$$

with $k(t) = \beta t^\theta$ and $d = \frac{\beta\lambda}{\alpha}$, where $\Gamma[\alpha, z_0, z_1] = \int_{z_0}^{z_1} t^{\alpha-1}e^{-t}dt$ denotes the generalized incomplete Gamma-function. For this model it is possible to show $v^{KM}(t) \geq v(t)$.

Proposition 5 *For all choices of parameters $\alpha, \beta, \lambda, \theta > 0$ and $\forall\ 0 \leq t \leq \tau^\star$ we have*

$$\begin{aligned} v^{KM}(t) - v(t) &= \lambda\left(\frac{(\alpha-\beta\lambda)(e^{k(t)}-1) - \alpha k(t)e^{k(t)}}{\beta(\alpha t^\theta+\lambda)^2} + \frac{\beta e^{(2k(t)+d)}\Gamma[t]}{\alpha}\right) \\ &\geq 0, \end{aligned} \tag{12}$$

where $\Gamma[t]$ is defined as in (11).

Proof.
To prove the proposition we will use Jensen's inequality in the following form (see e.g. Klambauer (1975)):

$$f\left(\frac{\int_a^b p(x)g(x)dx}{\int_a^b p(x)dx}\right) \leq \frac{\int_a^b p(x)f(g(x))dx}{\int_a^b p(x)dx},$$

where $f(\cdot)$ is a convex function on an interval (c_1, c_2), $g(\cdot)$ an integrable function on $[a, b]$ with $c_1 < g(x) < c_2$, the function $p(\cdot)$ is strictly positive on $[a, b]$ and all integrals exist.
In view of (12) we have to show that for all $\alpha, \beta, \lambda, \theta > 0$ and $\forall\ 0 \leq t \leq \tau^\star$:

$$\Gamma[0, k(t)+d, 2k(t)+d] \geq \frac{\alpha^2 k(t)e^{k(t)} - \alpha(\alpha-\beta\lambda)(e^{k(t)}-1)}{(\alpha k(t)+\beta\lambda)^2} \cdot e^{-(2k(t)+d)}$$

In this inequality we replace the left-hand side $\Gamma[t]$ by a lower bound which is derived using the Jensen-type inequality above for the special choice $a = k(t)+d$, $b = 2k(t)+d$, $c_1 = 0$, $c_2 = \infty$, $p(x) = \exp\{-x\}$, $f(x) = \frac{1}{x}$ and $g(x) = x$. Since all assumptions are fulfilled some algebra yields

$$\Gamma[0, k(t)+d, 2k(t)+d] \geq \frac{e^{-(k(t)+d)}\left(1-e^{-k(t)}\right)^2}{k(t)+d+1-e^{-k(t)}\left(2k(t)+d+1\right)}.$$

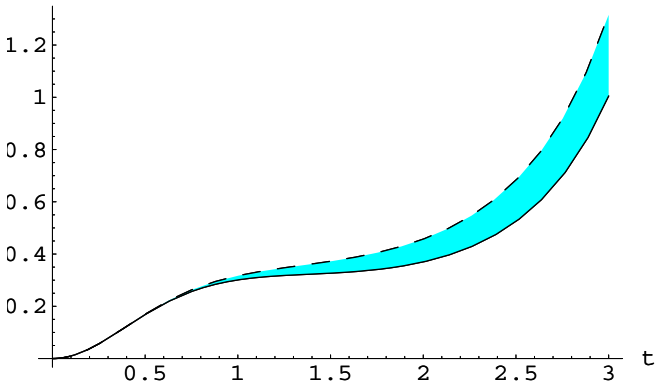

Figure 5: Asymptotic variance functions $v(t)$ —— and $v^{KM}(t)$ - - - for a parameter choice: $\alpha = 1$, $\beta = 0.457$, $\lambda = 1$ and $\theta = 2$.

The next step is to show that the right-hand side of this inequality above is an upper bound of $\frac{\alpha^2 k(t) e^{k(t)} - \alpha(\alpha - \beta\lambda)(e^{k(t)} - 1)}{(\alpha k(t) + \beta\lambda)^2} \cdot e^{-(2k(t)+d)}$, i.e.

$$\frac{e^{-(k(t)+d)} \left(1 - e^{-k(t)}\right)^2}{k(t) + d + 1 - e^{-k(t)} \left(2k(t) + d + 1\right)} \geq \frac{k(t)e^{k(t)} - (1-d)(e^{k(t)} - 1)}{(d + k(t))^2 e^{(2k(t)+d)}}.$$

After some lengthy, but straightforward calculations it turns out that this is equivalent to

$$e^{-k(t)} \left(k(t) + 1 - e^{k(t)}\right)^2 \geq 0,$$

what now easily verifies (12) for all $\alpha, \beta, \lambda, \theta > 0$ and $\forall\ 0 \leq t \leq \tau^\star$. ■

It seems that in general explicit expression for the asymptotic variance can hardly be determined. For a special setting of parameters ($\alpha = 1$, $\beta = 0.457$, $\lambda = 1$ and $\theta = 2$ with $P(\delta = 0) = 0.44$) a significant reduction of asymptotic variance especially in the tail can be observed (see Figure 5). This is something one might have expected given additional information provided in the model by the covariate Z. The behaviour of the asymptotic variance strongly depends on the choice of the rates α, β, λ and the shape parameter θ.

3.3 Leukemia Study (University of Ulm 1993)

Finally we present the analysis of data from a leukemia study carried out at the University of Ulm in 1993. In this study 80 persons with diagnosis of leukemia were examined

from the date of diagnosis till death, recovery or other reasons which took them off the study. A group of 39 persons was treated with a placebo drug while the remaining 41 persons received a leukemia drug. The covariate relapse-free survival time has been observed in addition to survival and censoring time and it seems plausible, that this covariate has effect on them. Since we aim at an estimation of the survival curve for the population within the two groups to be able to predict survival probabilities for patients with the same diagnosis, and of course the covariate is not known in advance, the CIM-model seems appropriate. We compute estimators for the survival function $\bar{F}(t)$ using the KM-estimator $\hat{\bar{F}}_n^{KM}(t)$ and the CIM-estimator $\hat{\bar{F}}_n(t)$ with the covariate relapse-free survival time. The examined data was of the form

$$(X_j, \delta_j, Z_j, \rho_j) \qquad j = 1, \ldots, 80 \qquad \text{with}$$

$$X_j = \min(T_j, U_j), \qquad \delta_j = \begin{cases} 0 \text{ if } T_j > U_j \\ 1 \text{ if } T_j \leq U_j \end{cases}, \qquad \rho_j = \begin{cases} 0 \text{ if} & \text{Placebo-Group} \\ 1 \text{ if} & \text{Verum- Group} \end{cases}$$

The following short extraction gives a little insight to the data:

X_j	δ_j	Z_j	ρ_j
4.3	0	2.2	1
5.2	1	1.7	1
5.4	0	2.5	0
6	0	2.6	1
6.1	0	3.8	1
⋮	⋮	⋮	⋮

The verum-group has a censoring proportion of 53.7% and the placebo-group of 58.9%. Figures 6 and 7 show the estimates stratified by treatment.
We conclude from Figure 6, that if survival and censoring times are independent or conditionally independent given the covariate Z, the estimators $\hat{\bar{F}}_n^{KM}(t)$ and $\hat{\bar{F}}_n(t)$ nearly reflect the same survival structure for the verum-group. For the placebo-data (Figure 7) a significant difference between $\hat{\bar{F}}_n^{KM}(t)$ and $\hat{\bar{F}}_n(t)$ can be observed. This difference is due to the additional information taken into account by using the covariate data.

Acknowledgement: We would like to thank Prof. F. Liese for drawing our attention to Cheng's paper and for helpful comments. We also thank the referees and an associate editor for their careful reading and suggestions which helped to improve the presentation of the paper.

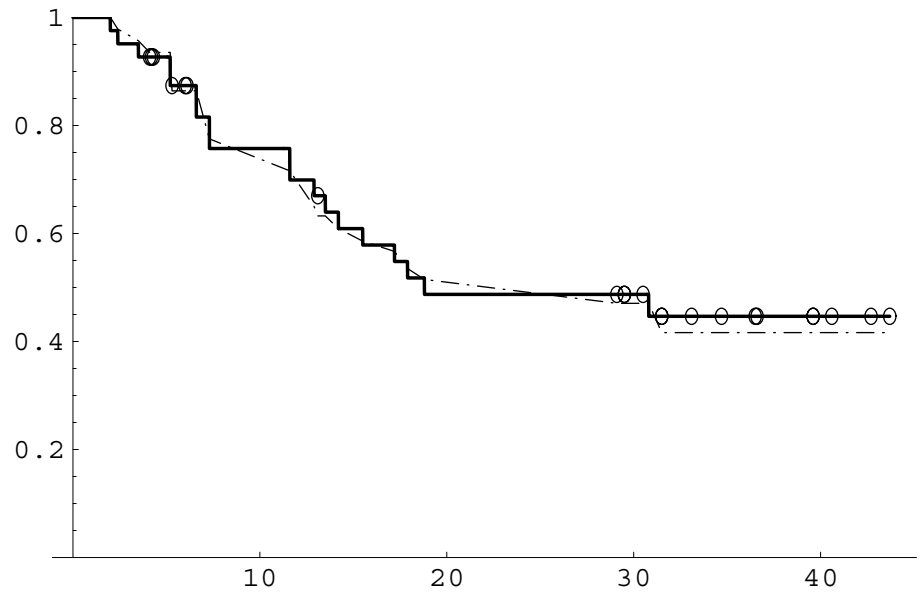

Figure 6: $\widehat{\bar{F}}_n^{KM}(t)$ —— and $\widehat{\bar{F}}_n(t)$ $-\cdot-$ for Verum-data ($\circ$ = censored observation)

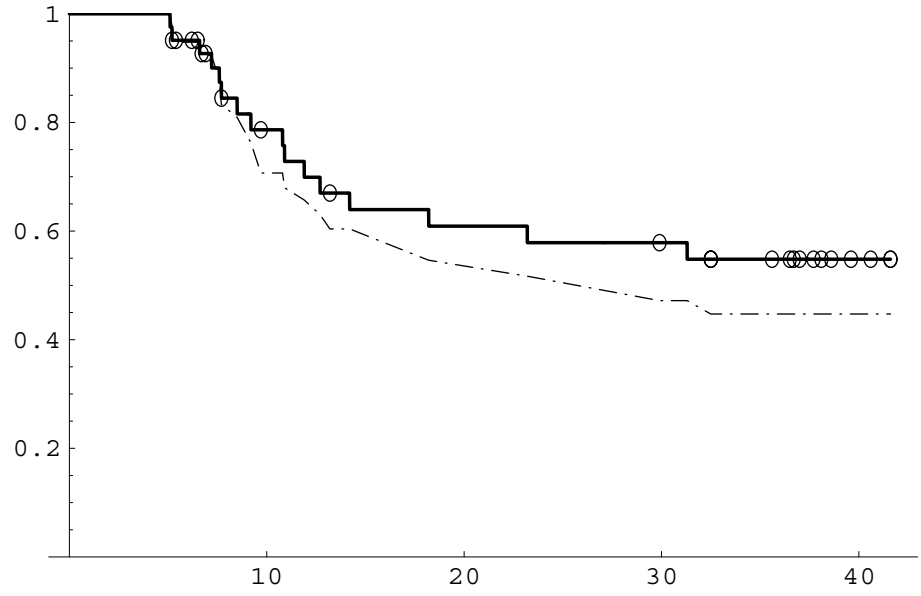

Figure 7: $\widehat{\bar{F}}_n^{KM}(t)$ —— and $\widehat{\bar{F}}_n(t)$ $-\cdot-$ for Placebo-data ($\circ$ = censored observation)

References

[1] Beran R (1981) Nonparametric regression with randomly censored survival data. Tech. Rep., University of California, Berkeley, CA

[2] Billingsley P (1968) Weak Convergence of Probability Measures. John Wiley&Sons, New York

[3] Breslow N, Crowley J (1974) A large sample study of the life table and product limit estimates under random censorship. Ann. Statist. 2:437-453

[4] Chen K, Lo S-H (1997) On the rate of uniform strong convergence of the product-limit estimator: strong and weak laws. Ann. Statist. 25:1050-1087

[5] Cheng PE (1989) Nonparametric estimation of survival curve under dependent censorship. J. Statist. Plann. Inference 23:181-191

[6] Csörgö S (1988) Estimation in the proportional hazards model of random censorship. Statistics 19, 437-463

[7] Csörgö S (1996) Universal Gaussian approximations under random censorship. Ann. Statist. 24, 2744-2778

[8] Csörgö S (1998) Testing for the partial proportional hazards model of random censorship. In: Prague Stochastics 98 Proceedings, Union of Czech Mathematicians and Physicists, Prague. 87-92

[9] Csörgö S, Faraway JJ (1998) The paradoxical nature of the proportional hazards model of random censorship. Statistics 31, 67-78.

[10] Dabrowska DM (1987) Nonparametric regression with censored survival time data. Scand. J. Statist. 14:181-197

[11] Dabrowska DM (1989) Uniform consistency of the kernel conditional Kaplan-Meier estimate. Ann. Statist. 17:1157-1167

[12] Dikta G (1998) On semiparametric random censorship models. J. Statist. Plann. Inference 66:253 - 279

[13] Gather U, Pawlitschko J (1998) Estimating the survival function under a generalized Koziol-Green model with partially informative censoring. Metrika 48:189-207

[14] Jensen U, Wiedman J (2000) Estimating a survival function from censored data in the presence of a discrete covariate. Research report, University of Ulm

[15] Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53:457-481

[16] Klambauer G (1975) Mathematical analysis. Marcel Dekker, Inc., New York

[17] Lo S-H, Singh K (1985) The product-limit estimator and the bootstrap: some asymptotic representations. Probab. Th. Rel. Fields 71:455-465

[18] Langberg N, Proschan F, Quinzi AJ (1978) Converting dependent models into independent ones, preserving essential features. Ann. Probab. 6:174-181

[19] Nair VN (1993) Bounds for reliability estimation under dependent censoring. Int. Statist. Rev. 61:169-182

[20] Rosenblatt M (1971) Curve estimates. Ann. Math. Statist. 42:1815-1842

[21] Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. John Wiley & Sons, New York

[22] Stute W, Wang J-L (1993) The strong law under random censorship. Ann. Statist. 21:1591-1607

[23] Williams JS, Lagakos SW (1977) Models for censored survival analysis: Constant-sum and variable-sum models. Biometrika 64:215-224