Some simulation results under random censorship models

by

Gerhard Dikta, Rolf Hansmann, and Christian Schmidt

Fachhochschule Aachen, Abt. Jülich

Abstract. We consider the Kaplan-Meier and a new semiparametric approach to estimate the distribution function in the random censorship model. Corresponding estimators of several functionals are compared by Monte Carlo under heavy censoring. Furthermore, some simulation results are included to illustrate the influence of linear dependence between lifetime and censoring time on the different approaches.

AMS 1991 subject classification. Primary 62E25; secondary 62F10, 62A10, 62G30.

Key words and phrases. Censored data, maximum likelihood estimation, uniform consistency, Kaplan-Meier estimator, proportional hazards model.

1. Introduction. In the statistical analysis of failure time or, a bit more positive, lifetime data much attention is paid to handle incomplete observations. Incompleteness in this scenario is mainly caused by censoring.

One important type of censoring is described by the Random Censorship Model (RCM). This model is widely accepted in practice and often the basic underlying assumption in theoretical considerations and practical data analysis. For this assume that $X_1, ..., X_n$ are independent and identically distributed (i.i.d.) positive random variables, the lifetimes, which are defined on some probability space $(\Omega, \mathcal{A}, I\!\!P)$, with unknown distribution function (df.) F about which some inference has to be made. In the RCM these data are censored on the right and one observes

$$Z_i = min(X_i, Y_i) \quad \text{and} \quad \delta_i = 1_{[X_i \leq Y_i]}, \qquad 1 \leq i \leq n,$$

where $Y_1, ..., Y_n$ are the censoring times, i.e. another sequence of positive i.i.d. random variables with df. G which are also independent of the X's. The variable δ_i indicates whether X_i is censored $(\delta_i = 0)$ or not $(\delta_i = 1)$. In this text we denote the df. of Z by H. For some practical applications of the RCM in relevant examples the reader is reffered to Andersen et al (1993).

To estimate some functional $T(F)$ based on the observations $(Z_1, \delta_1), \cdots, (Z_n, \delta_n)$, one naturally tries to use $T(\hat{F}_n)$, where $\hat{F}_n$ is an estimator of F. Obviously, a reasonable estimator of F should incorporate all observations, also the censored ones. In this article we compare $T(\hat{F}_n)$ for three different estimators of F by Monte-Carlo. These estimators are especially designed for applications under RCM. In particular, we consider the nonparametric Kaplan-Meier estimator, a parametric estimator based on a modified maximum likelihood approach, and a new semiparametric estimator. Our attention here is mainly focused on the performance of the latter one compared to the Kaplan-Meier estimator.

The aforementioned first candidate for $\hat{F}_n$ is the nonparametric estimator of F, the time-honored Kaplan-Meier (1958) product limit estimator defined by

$$1 - F_n^{km}(t) = \prod_{i:Z_i \leq t} \left(1 - \frac{\delta_i}{n - R_i + 1}\right) \tag{1}$$

Here R_i denotes the rank of Z_i within the Z-sample. The Kaplan-Meier estimator is widely used in practice, since it is implemented in almost every statistical software package, and received great attention in the literature. Uniform consistency of the Kaplan-Meier estimator was shown by Shorack and Wellner, see Shorack and Wellner (1986 p. 304). In particular, they proved that

$$\| F_n^{km} - F \|_0^{H^{-1}(1)} = \sup_{0 \leq t < H^{-1}(1)} | F_n^{km}(t) - F(t) |,$$

tends to 0 with probability one. Here $H^{-1}(u) = \inf\{t : H(t) \geq u\}$, with $0 \leq u \leq 1$, defines the u quantile of H. Breslow and Crowley (1974) investigated the Kaplan-Meier process

$$n^{1/2}(F_n^{km}(t) - F(t)), \quad \text{for} \quad 0 \leq t \leq T,$$

with $H(T) < 1$ and proved a functional central limit theorem. A corresponding result for the quantile process was given by Sander (1975), see Shorack and Wellner (1986 p. 657). Some of the functionals which are considered here can be expressed by integrals with respect to F. General results on strong consistency and asymptotic normality of Kaplan-Meier integrals over the whole real line can be found in Stute and Wang (1993) and Stute (1995), respectively. Furthermore, sharp bounds for the bias of Kaplan-Meier integrals are given in Stute (1994).

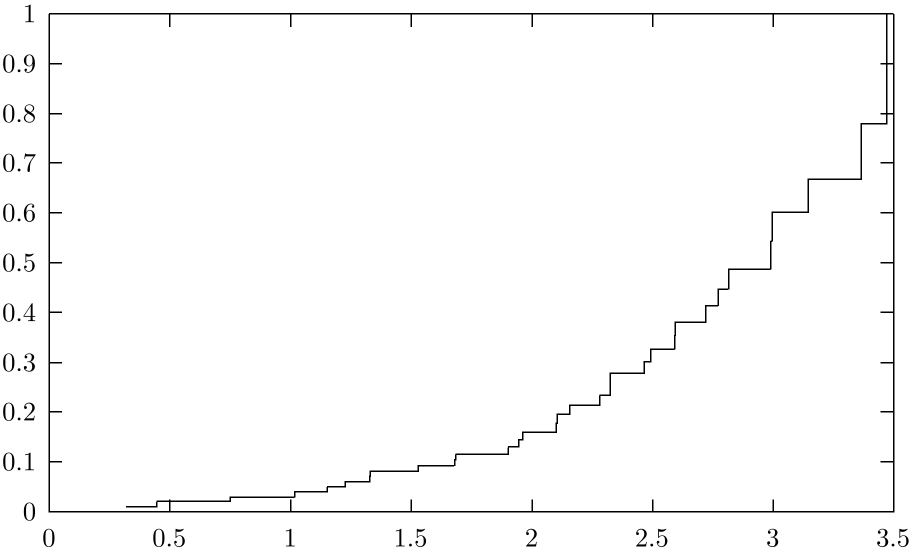

Figure 1: The figure shows the Kaplan-Meier estimator F_{100}^{km} based on 100 simulated data with 68 observations being censored.

Note, that F_n^{km} attaches mass only to the uncensored observations. Efron (1967) pointed out that the attached mass increases from the smallest to the largest uncensored observation, where the amount of increase between two uncensored observations depends on the number of censored observations between them. Therefore, under heavy censoring F_n^{km} has only a few jumps with increasing sizes (see figure 1) and a practitioner might not be satisfied with the accuracy of F_n^{km}.

If we have good reasons to assume that F has a density f which belongs to a known parametric family,

$$f(t) \quad = \quad f(t, \gamma_0),$$

where $\gamma_0 = (\gamma_{0,1}, ..., \gamma_{0,j}) \in \Gamma$, a maximum likelihood approach is appropriate and the unknown parameter γ_0 is estimated by the MLE $\hat{\gamma}_n$, i.e. $\hat{\gamma}_n$ is the maximizer of the likelihood function

$$L_n^{pa}(\gamma) \quad = \quad \prod_{i=1}^{n} \left[f(Z_i, \gamma)^{\delta_i} (1 - F(Z_i, \gamma))^{1-\delta_i} \right]. \tag{2}$$

The estimated df. according to the parametric approach is denoted here by F_n^{pa}, where F_n^{pa} abbreviates $F(\cdot, \hat{\gamma}_n)$. For the parametric estimation under RCM an extensive literature is available. Some general references can be found in Kalbfleisch and Prentice (1980), where also asymptotic normality of the MLE is discussed. For general results on strong consistency the reader is referred to Stute (1992b).

The parametric approach leads to an estimator which is sufficiently accurate in practice. However, if we are not sure about the parametric family for the underlying density, this approach can not be used.

Our third candidate to estimate F is based on a semiparametric approach, which is a compromise between the approaches just discussed. The estimator is defined by

$$1 - F_n^{se}(t) \quad = \quad \prod_{i: Z_i \le t} \left(1 - \frac{m(Z_i, \hat{\theta}_n)}{n - R_i + 1} \right), \tag{3}$$

where $m(t)$ denotes the conditional expectation of δ given $Z = t$, i.e.

$$m(t) \; = \; I\!P(\delta = 1 \mid Z = t) \; = \; I\!E(\delta \mid Z = t).$$

Under this approach it is assumed that m belongs to a parametric family so that we can write

$$m(t) \quad = \quad m(t, \theta_0),$$

where $m(\cdot, \cdot)$ is a known function and $\theta_0 = (\theta_{0,1}, ..., \theta_{0,k}) \in \Theta$. The unknown parameter θ_0 is estimated by the MLE $\hat{\theta}_n$, i.e. the maximizer of the likelihood function

$$L_n^{se}(\theta) \quad = \quad \prod_{i=1}^{n} \left[m(Z_i, \theta)^{\delta_i} \cdot (1 - m(Z_i, \theta))^{1-\delta_i} \right]. \tag{4}$$

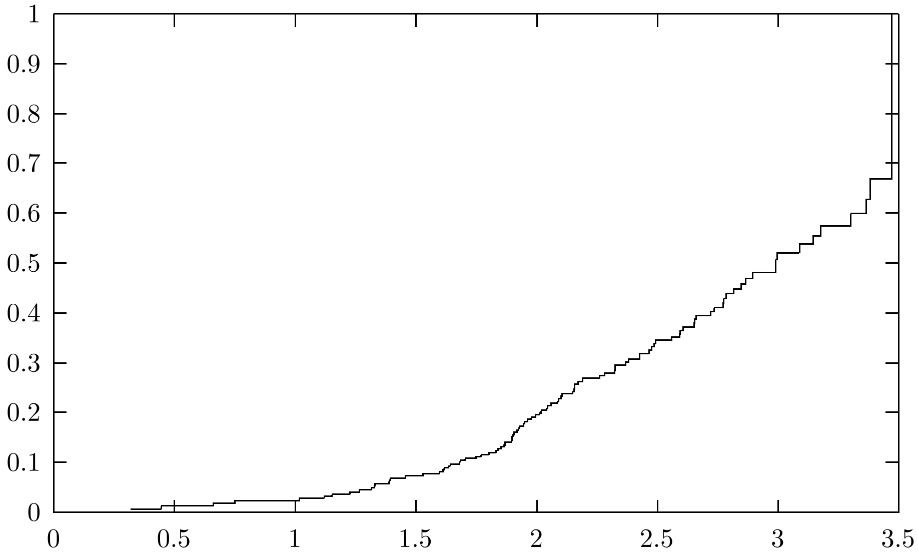

Figure 2: The figure shows the semiparametric estimator F_{100}^{se} based on the same 100 simulated data used for F_{100}^{km} in figure 1.

The semiparametric estimator was proposed by Dikta (1998) who stated uniform consistency, a strong law for F_n^{se} integrals over the whole real line, Dikta (2000), and a functional central limit result for the corresponding process, where F_n^{se} was studied on some compact interval $[0, T]$ with $H(T) < 1$. Comparing the covariance structure of the limiting process with the one given by Breslow and Crowley (1974) for the Kaplan-Meier process, a gain in efficiency for F_n^{se} towards F_n^{km} in terms of asymptotic variance was obtained. Parametric forms for m were motivated by the equation

$$m(t) \quad = \quad \frac{\lambda_f(t)}{\lambda_h(t)} \quad = \quad \frac{\lambda_f(t)}{\lambda_f(t) + \lambda_g(t)}, \tag{5}$$

where f, g, h denote the density functions of X, Y, Z and $\lambda_f, \lambda_g, \lambda_h$ the corresponding hazard functions, respectively. Furthermore, it was pointed out that the parametric family $m(t, \theta) = \theta$ specifies the Proportional Hazards Model (PHM), see Koziol-Green (1976), and that F_n^{se} is under PHM identical to F_n^{cl}, the Cheng and Lin (1987) estimator, see also Abdushukurov (1987). Thus, F_n^{se} generalizes F_n^{cl}. See Csörgő (1988) for a review of fundamental properties of F_n^{cl}. In the special case of the PHM results on strong consistency and asymptotic normality of integrals over the whole real line with respect to F_n^{cl} can be found in Stute (1992a) and Dikta (1995), respectively. More recently, the stated semiparametric approach was applied by Sun and Zhu (2000) for truncated data.

If we compare figure 1 with figure 2 we get the impression that F_n^{se} is more accurate than F_n^{km}. This is indeed the case since F_n^{se} attaches weights onto all the observations and not only to the uncensored ones, as F_n^{km} does. Therefore, there are more jumps and they do not necessarily increase when we use

F_n^{se}. To apply the parametric approach one has to restrict to a certain parametric class of densities. In the semiparametric approach there is also a restriction to a parametric class for m. But, as (5) shows, one can choose the class for m in such a way that it contains the class of densities and even more than just these densities. Thus, the semiparametric approach is not as restrictive as the parametric one. Nevertheless, there are restrictions when the semiparametric approach is used which are not necessary for the Kaplan-Meier estimator. Finally, diagnostics to check the assumed parametric model m of F_n^{se} can be based on methods used in the analysis of binary data, see Cox and Snell (1989). In particular, we assume here a model for the binary data $\delta_1, \delta_2, \ldots, \delta_n$ with conditional probability of success given by $m(Z_1), m(Z_2), \ldots, m(Z_n)$.

The objective of the present paper is to compare estimators based on F_n^{km} with those derived from F_n^{se} for a moderate sample size under heavy censoring in a simulation study.

Since asymptotically F_n^{se} is more efficient than the Kaplan-Meier estimator, if the semiparametric model is correctly specified, it is important for practical data analysis to have some results about the performance of both estimators under finite sample sizes when the semiparametric model is correctly and incorrectly specified.

Finally, in clinical studies, where censoring is caused by lost to follow-up cases, the general assumption of independence between X and Y might fail to hold because the disappearance of a patient from the study might be caused by the treatment itself. In this case, one of the general assumptions of the RCM is violated and we are interested in the effect of this violation on the different estimation procedures. Therefore, we included here some simulations to analyze the influence of linear dependence between X and Y on the different approaches.

Due to the improvement of technological items and further developments in the medical field we will have to handle more often heavily censored data sets in future, since increasing lifetimes lead to an increase in censoring, if the testing time will not be enlarged. But an enlargement of testing time is definitely unacceptable in the technological or medical field. Therefore, reasonable competitors to the Kaplan-Meier estimator will be needed in data analysis to handle these type of data sets. As already pointed out, a pure parametric approach might not be possible in this scenario.

2. General computational aspects. F_n^{km} and F_n^{se} are step functions which attach mass only to the observed Z-values. In particular, the weight attached to $Z_{(i)}$ based on the Kaplan-Meier estimator is given by

$$W_{i,n}^{km} = F_n^{km}(Z_{(i)}) - F_n^{km}(Z_{(i-1)}) = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{\delta_{(j)}}{n-j+1}\right), \qquad 1 \le i < n.$$

Here $(Z_{(1)}, \delta_{(1)}), \cdots, (Z_{(n)}, \delta_{(n)})$ denotes the ordered $Z-$sample with the associated δ's and $Z_{(0)} \equiv 0$. Furthermore, an empty product is always defined as 1. Since $F_n^{km}(Z_{(n)})$ is less than 1 if the last datum is censored, we set $F_n^{km}(Z_{(n)}) \equiv 1$ to have a real df. and therefore the weight w.r.t. the Kaplan-Meier estimator for the largest observation is given by

$$W_{n,n}^{km} = 1 - F_n^{km}(Z_{(n-1)}) = \prod_{j=1}^{n-1} \left(1 - \frac{\delta_{(j)}}{n-j+1}\right).$$

Expectation and variance based on F_n^{km} are then obtained by

$$\begin{aligned} I\!E_n^{km}(X) &= \sum_{i=1}^{n} Z_{(i)} W_{i,n}^{km} \\ VAR_n^{km}(X) &= \sum_{i=1}^{n} (Z_{(i)} - I\!E_n^{km}(X))^2 W_{i,n}^{km}. \end{aligned}$$

The quantiles are estimated by the corresponding quantiles of F_n^{km}, i.e.

$$(F_n^{km})^{-1}(u) \quad = \quad \inf\{t : F_n^{km}(t) \geq u\}, \qquad 0 \leq u \leq 1.$$

The conditional expectation based on F_n^{km} was calculated by

$$\begin{aligned} & I\!E_n^{km}(X \mid (F^{-1}(0.25) < X \leq F^{-1}(0.75)) \\ = \; & \frac{1}{F_n^{km}((F_n^{km})^{-1}(0.75)) - F_n^{km}((F_n^{km})^{-1}(0.25))} \sum_{i:\, 0.25 < F_n^{km}(Z_{(i)}) \leq 0.75} Z_{(i)} W_{i,n}^{km} \end{aligned}$$

To measure how accurate F is estimated by the considered procedures, one can take the uniform distance between the estimated df. and F over the whole real line. However, for both F_n^{km} and F_n^{se} the uniform distance over the whole real line is mainly determined by the last observation. Therefore, the uniform distance in our simulation was calculated only up to $Z_{(n-1)}$, e.g. for F_n^{km} by

$$\sup_{0 \leq t \leq Z_{(n-1)}} \mid F_n^{km}(t) - F(t) \mid .$$

Another measure of the accuracy of the estimated df. is the L_2 distance. This was calculated, again, only up to $Z_{(n-1)}$, e.g. for F_n^{km} by

$$\int_0^{Z_{(n-1)}} \left(F_n^{km}(t) - F(t)\right)^2 dt.$$

To calculate the weights based on the semiparametric estimator, one first has to choose an appropriate parametric model for m. Next, $\hat{\theta}_n$ the MLE corresponding to (4) has to be obtained. The weights are then given by

$$W_{i,n}^{se} \;=\; F_n^{se}(Z_{(i)}) - F_n^{se}(Z_{(i-1)}) \;=\; \frac{m(Z_{(i)}, \hat{\theta}_n)}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \frac{m(Z_{(j)}, \hat{\theta}_n)}{n-j+1}\right), \qquad 1 \leq i < n.$$

As for the Kaplan-Meier estimator we set here $F_n^{se}(Z_{(n)}) \equiv 1$ in order to get a real df. and therefore, the largest observation is weighted by

$$W_{n,n}^{se} \;=\; 1 - F_n^{se}(Z_{(n-1)}) \;=\; \prod_{j=1}^{n-1} \left(1 - \frac{m(Z_{(j)}, \hat{\theta}_n)}{n-j+1}\right), \qquad 1 \leq i < n.$$

Note that $W_{i,n}^{km} = 0$ if $Z_{(i)}$ is censored, while the corresponding weight based on F_n^{se} is positive, at least for a reasonable choice of m. The estimators based on F_n^{se} were defined similarly to those corresponding to F_n^{km}, i.e. take the same formulae and replace the weights.

In the parametric approach one first has to choose a parametric family for the density f. With $\hat{\gamma}_n$, the MLE corresponding to (2), the values of interest can then be calculated with respect to the estimated density $f(\cdot, \hat{\gamma}_n)$. Needless to say, the distances under the parametric approach were also calculated from 0 up to $Z_{(n-1)}$.

All computations have been performed on a PC using C and some routines of the Numerical Recipes in C, see Press et al. (1992), under LINUX.

3. Results of the simulation study. In our study we chose a two-parameter Weibull distribution for the lifetime df. F, i.e.

$$f(t) \quad = \quad \alpha\beta(\alpha t)^{\beta-1} \exp(-(\alpha t)^{\beta}), \qquad t \geq 0,$$

with $\alpha = 0.3$ and $\beta = 3$, since this is the classical type of df. in survival analysis.

The values of interest based on F were calculated numerically to $I\!E(X) \approx 2.977$, $VAR(X) \approx 1.170$, $F^{-1}(0.25) \approx 2.200$, $F^{-1}(0.50) \approx 2.950$, $F^{-1}(0.75) \approx 3.717$, and $I\!E(X \mid F^{-1}(0.25) < X \leq F^{-1}(0.75)) \approx 2.953$.

The results of our simulations are based for each model on 1000 replications of data sets with sample size $n = 100$. For each estimated value we subtracted the true value and took the average over the 1000 replications in order to estimate the bias. Furthermore, for each estimator the variance over the replications were taken to estimate the variance of the estimator. The results are illustrated in the tables below, denoted there by $\widehat{BIAS}$ and $\widehat{VAR}$, respectively. Finally, we denote by $\widehat{AVE}$ the average over the 1000 replications of the uniform and of the L_2 distance, respectively.

Model 1: In the first model we censored F with the Weibull distribution G, where $\alpha = 0.4$ and $\beta = 4$. Under this model about 70% of the observations are censored.

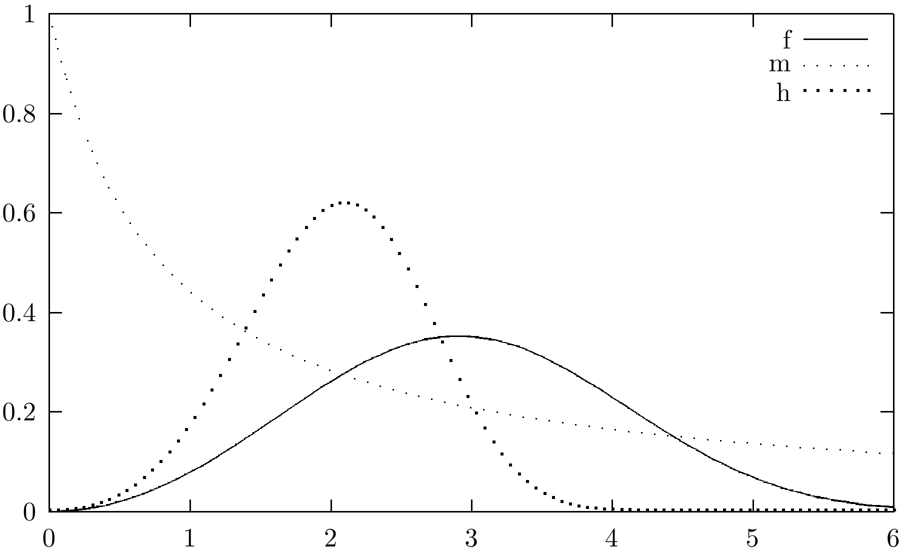

Figure 3: The figure shows the true density f, the conditional expectation of δ given $Z = t$ (i.e. m), and the density h of the observations Z for Model 1.

Figure 3 illustrates that especially the larger observations of Z are censored. Therefore, $F_n^{km}(Z_{(n-1)})$ can be expected to be substantially less than 1. Furthermore, we can hardly expect to have observations Z above $F^{-1}(0.75)$.

For the parametric procedure the Weibull family was assumed. According to (5) the parametric form for m in the case where F and G are both Weibull distributions, is given by

$$m(t, \theta) \quad = \quad \frac{\theta_1}{\theta_1 + t^{\theta_2}}, \tag{6}$$

where $\theta_1 > 0$ and $\theta_2 \in I\!R$, see Example 2.9 in Dikta (1998). We chose this two-parameter model for the semiparametric estimator since it can handle monotonic m's.

Note that the parametric model assumptions are taken properly. Therefore, the parametric approach should produce the best estimates, while the results based on the semiparametric one should be better than those derived from the Kaplan-Meier approach.

	KM		SE		PA	
	$\widehat{BIAS}$	$\widehat{VAR}$	$\widehat{BIAS}$	$\widehat{VAR}$	$\widehat{BIAS}$	$\widehat{VAR}$
expectation	-0.212	0.020	-0.206	0.018	0.031	0.058
variance	-0.543	0.026	-0.539	0.025	0.091	0.296
cond. expect.	-0.252	0.035	-0.161	0.012	0.025	0.048
$F^{-1}(0.25)$	0.014	0.031	0.013	0.022	0.015	0.019
$F^{-1}(0.50)$	0.072	0.093	0.058	0.060	0.024	0.046
$F^{-1}(0.75)$	-0.252	0.054	-0.212	0.047	0.039	0.133

Table 1: Estimated expectation, variance, conditional expectation, and quantiles under Model 1.

Table 1 shows what we expected. The best results are achieved by the pure parametric approach, while the estimates of expectation and variance show almost the same quality under the semiparametric and the nonparametric approach. However, the semiparametric estimate of the conditional expectation is remarkably better than the nonparametric one. Nevertheless, under the semiparametric and the nonparametric approach the expectation, the variance, and the conditional expectation are underestimated, as the negative bias shows. Note, that in many samples $F_n^{km}(Z_{(n-1)}) < 0.75$ and $F_n^{se}(Z_{(n-1)}) < 0.75$. Thus $Z_{(n)}$ was taken to estimate the 75%-quantile of F under both approaches in these cases. But, as figure 3 shows, we can hardly expect an observation of Z to be above $F^{-1}(0.75)$. Therefore, the semiparametric and the nonparametric approach underestimate the 75%-quantile of F. Overall, the semiparametric estimates are better than the nonparametric ones and both are outperformed by the estimates based on the pure parametric approach.

	KM		SE		PA	
	$\widehat{AVE}$	$\widehat{VAR}$	$\widehat{AVE}$	$\widehat{VAR}$	$\widehat{AVE}$	$\widehat{VAR}$
uniform distance	0.1378	0.0027	0.0676	0.0006	0.0030	0.0000
L_2 distance	0.0045	0.0125	0.0011	0.0007	0.0054	0.0477

Table 2: Uniform and L_2 distances under Model 1.

Table 2 shows for the uniform distance what could be expected, i.e. the parametric approach is the most accurate one, while the semiparametric approach leads to better estimates of F than the Kaplan-Meier approach, at least for our modified uniform distance. However, w.r.t. the L_2 distance the parametric estimates are outperformed by the two other candidates which is indeed remarkable. Overall, the semiparametric estimates of F w.r.t. the L_2 distance give the best results.

To get some impression of the performance, the next figures show the semiparametric and the Kaplan-Meier estimates of two data sets. In the first figure we choose a data set which favors the semiparametric estimator w.r.t. the L_2 distance.

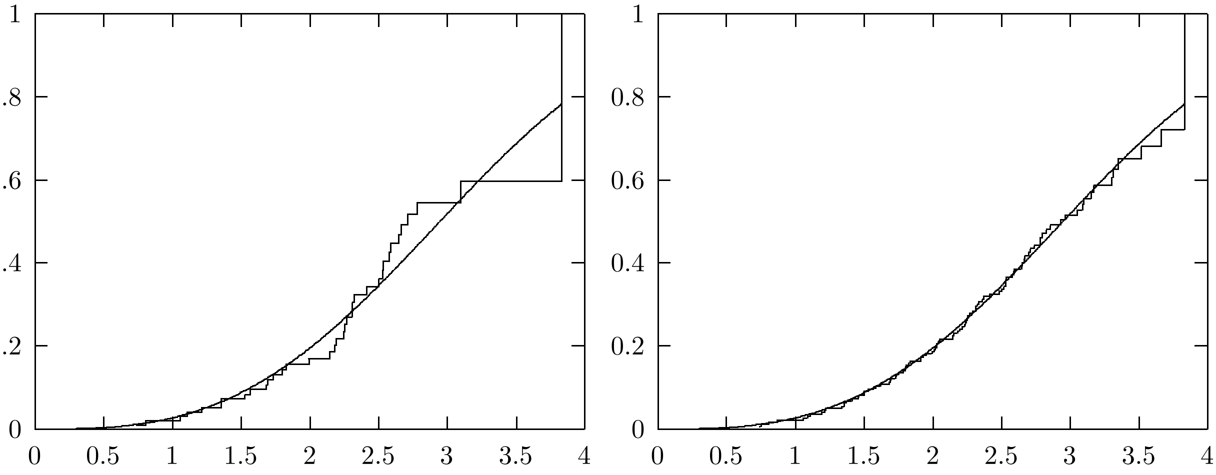

Figure 4: The figure shows on the right-hand side the true df. F and F_n^{se} for one of the simulated data sets under Model 1, which favors F_n^{se} w.r.t. the L_2 distance. The left-hand side shows F and F_n^{km} for the same data set.

The calculated L_2 distance for the semiparametric estimator for this data set is 0.0004 and for the Kaplan-Meier estimator 0.0052. Obviously, F_n^{se} fits F very well while F_n^{km} loses accuracy due to some under- and overestimated regions.

The chosen data set of the next figure favors the Kaplan-Meier estimator w.r.t. the L_2 distance

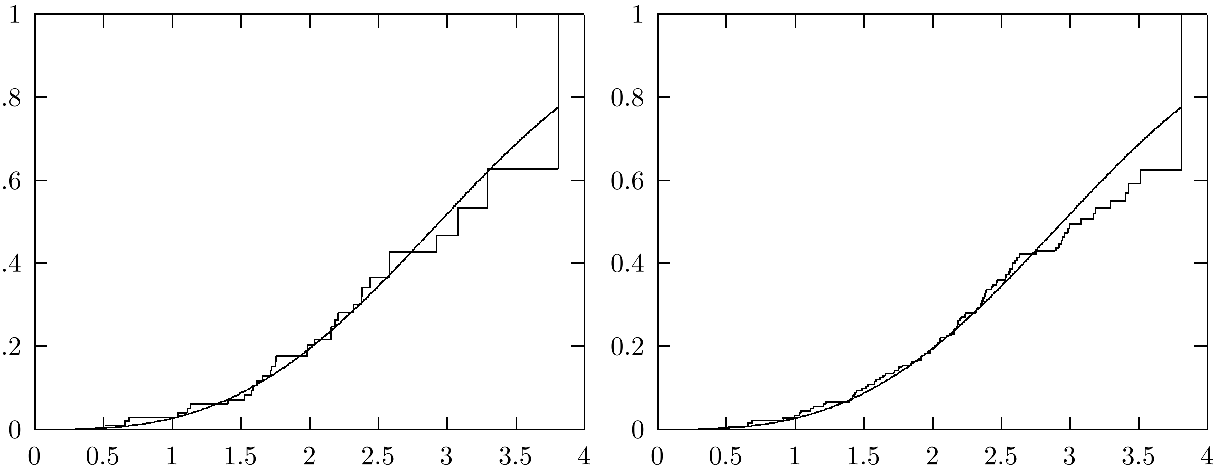

Figure 5: The figure shows on the left-hand side the true df. F and F_n^{km} for one of the simulated data sets under Model 1, which favors F_n^{km} w.r.t. the L_2 distance. The right-hand side shows F and F_n^{se} for the same data set.

The calculated L_2 distance for the Kaplan-Meier estimator for this data set is 0.0015 and for the semiparametric estimator 0.0022. Here, F_n^{se} loses accuracy due to the underestimation in the right tail of F, while the Kaplan-Meier estimator shows a good performance due to the fact that some of the larger observations are uncensored here.

In summary the simulations under model 1 show what was expected. A pure parametric estimator is the first choice if a parametric model can be assumed. Furthermore, the semiparametric estimator outperforms the Kaplan-Meier estimator under model 1 and the asymptotic gain in efficiency of F_n^{se} compared to F_n^{km} can also be observed under moderate sample size, provided that the correct parametric model for m is chosen. However, the high quality in accuracy of the semiparametric estimator w.r.t. the L_2 distance was neither expected nor could it be foreseen that the parametric estimator could be outperformed by the semiparametric one w.r.t. this measure of accuracy.

Model 2: Here, F was censored with a the uniform distribution on $[0, 4]$. Under this setup again about 70% of the observations are censored. For the parametric form of m again the one given under (6) was used. Note, that now the assumed semiparametric model is not correct anymore, while the assumptions for the Kaplan-Meier estimator are still satisfied. In this scenario we are only interested in the performance of the semiparametric estimates compared to those based on the Kaplan-Meier estimator, and therefore only the results of these two approaches are listed in the tables below.

	KM		SE	
	$\widehat{BIAS}$	$\widehat{VAR}$	$\widehat{BIAS}$	$\widehat{VAR}$
expectation	-0.143	0.018	-0.123	0.018
variance	-0.405	0.020	-0.420	0.016
cond. expect.	-0.120	0.018	0.015	0.018
$F^{-1}(0.25)$	0.002	0.034	0.084	0.032
$F^{-1}(0.50)$	0.015	0.061	0.081	0.045
$F^{-1}(0.75)$	-0.060	0.058	-0.103	0.041

Table 3: Estimated expectation, variance, conditional expectation, and quantiles under Model 2.

Surprisingly, the best results are achieved by the semiparametric approach here for the expectation, the variance, and the conditional expectation, even though a wrong model was used. The estimated quantiles based on the Kaplan-Meier approach however are, w.r.t. the bias significantly better than those based on the semiparametric approach.

	KM		SE	
	$\widehat{AVE}$	$\widehat{VAR}$	$\widehat{AVE}$	$\widehat{VAR}$
uniform distance	0.1636	0.0027	0.1449	0.0024
L_2 distance	0.0129	0.0001	0.0106	0.0001

Table 4: Uniform and L_2 distances under Model 2.

As under model 1 the semiparametric estimator shows a better performance than Kaplan-Meier estimator w.r.t. the uniform and the L_2 distance.

The performace of the estimators based on the incorrectly specified semiparametric model are similar to those derived by the Kaplan-Meier approach. However, the quantiles are better estimated by the latter approach. In summary, this indicates that the semiparametric estimations seem to be quite robust.

Model 3: The simulations under this model are intended to illustrate the effect of linear dependence between X and Y on the three approaches. We were interested in this effect since the independence assumption between X and Y in the RCM might be violated for a concrete data set. In particular, when there are lots of lost to follow-up cases in a clinical study we should be careful in assuming independence between the lifetime and the censoring time, especially since this assumption is untestable, see Kalbfleisch and Prentice (1980 pp. 172). The data sets under this model are based on

$$Y \quad = \quad 0.5 + 0.85 \cdot X + \varepsilon,$$

where ε is uniformly distributed on $[-0.25\,,\,0.25]$ and independently chosen from X. Under these assumptions about 40% of the observations are censored.

For the parametric approach the Weibull family was assumed and we used the two-parameter model given by (6) for the semiparametric one.

	KM		SE		PA	
	$\widehat{BIAS}$	$\widehat{VAR}$	$\widehat{BIAS}$	$\widehat{VAR}$	$\widehat{BIAS}$	$\widehat{VAR}$
expectation	0.440	0.039	0.448	0.038	0.510	0.045
variance	0.930	0.255	0.943	0.234	1.410	0.424
cond. expect.	0.228	0.027	0.296	0.031	0.421	0.038
$F^{-1}(0.25)$	0.049	0.027	0.056	0.025	0.100	0.021
$F^{-1}(0.50)$	0.289	0.053	0.263	0.048	0.406	0.037
$F^{-1}(0.75)$	1.221	0.405	1.266	0.301	0.809	0.095

Table 5: Estimated expectation, variance, conditional expectation, and quantiles under Model 3.

Note, the worst candidate to estimate expectation, variance, and conditional expactation is the parametric one. All three procedures overestimate the quantiles substantially. This indicates that the true df. is larger than any of the estimated ones, at least in the region where the data are observed.

	KM		SE		PA	
	$\widehat{AVE}$	$\widehat{VAR}$	$\widehat{AVE}$	$\widehat{VAR}$	$\widehat{AVE}$	$\widehat{VAR}$
uniform distance	0.1074	0.0012	0.0924	0.0011	0.0186	0.0002
L_2 distance	0.0075	0.0310	0.0067	0.0264	0.0600	1.4700

Table 6: Uniform and L_2 distances under Model 3.

The best result of the uniform distance based on the Kaplan-Meier approach under the 3 models is achieved in the third one, i.e. where the requirements of the RCM are not fulfilled. But, as the overestimated quantiles indicate the true df. lies above the estimated one. Therefore, the good result under model 3 for the uniform distance is misleading.

Overall, Kaplan-Meier and the semiparametric approach seem to be more robust against this violation of the RCM than the parametric approach.

4. Conclusion. The results of the simulations under model 1 confirm the gain in efficiency of the semiparametric approach compared to the nonparametric approach of Kaplan-Meier for the chosen sample size. Therefore, under heavy censoring, when the results obtained under the Kaplan-Meier estimator are not sufficient for a practitioner, estimators based on the considered semiparametric approach seem to be a reasonable alternative. Even under an incorrect model assumption the performance of the semiparametric estimators based on the two-parameter family considered here seems to be similar to the performance of the estimators which are based on the nonparametric approach of Kaplan-Meier.

Overall, the first choice is the parametric approach, if the correct family of densities can be specified and if independence between lifetime and censoring time is guaranteed. Otherwise, under an incorrect model assumption, or if the data do not fulfill the requirements of RCM, the results from this approach are misleading.

Finally, all three approaches fail when censoring is depending on the lifetime, i.e. RCM does not hold. However, F_n^{km} and F_n^{se} seem to be more robust against linear dependence between X and Y than the parametric candidate.

References

Abdushukurov, A.A., Nonparametric estimation in the proportional hazards model of random censorship. *Akad. Nauk Uz Tashkent.* **VINITI No. 3448-V** (1987) (in Russian).

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

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

Cheng, P.E. and G.D. Lin, Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model. *Statist. Probab. Letters* **5** (1987) 75-80.

Cox, D.R. and E.J. Snell, *Analysis of Binary Data, 2nd ed.* (Chapman Hall, London, 1989).

Csörgő, S., Estimation in the proportional hazards model of random censorship. *Statistics* **19** (1988) 437-463.

Dikta, G., Asymptotic normality under the Koziol-Green model. *Commun. Statist.-Theory Meth.* **24** (1995) 1537-1549.

Dikta, G., On semiparametric random censorship models. *J. Statist. Plann. Inference* **66** (1998) 253-279.

Dikta, G., The strong law under semiparametric random censorship models. *J. Statist. Plann. Inference* **83** (2000) 1-10.

Efron, B., The two-sample problem with censored data. *Proc. 5th Berkeley Symp.* **4** (1967) 831-853.

Kalbfleisch, J.D. and R.L. Prentice, *The Statistical Analysis of Failure Time Data.* (Wiley, New York, 1980).

Kaplan, E.L. and P. Meier, Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** (1958) 457-481.

Koziol, J.A. and S.B. Green, A Cramér-von Mises statistic for randomly censored data. *Biometrika* **63** (1976) 465-474.

Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C, 2nd ed.* (Cambridge University Press, 1992).

Sander, J.M., *The weak convergence of quantiles of the product.* Technical Report No. 5 (1975), Division of Biostatistics, Stanford University, Stanford, California.

Shorack, G.R. and J.A. Wellner, *Empirical Processes with Applications to Statistics.* (Wiley, New York, 1986).

Stute, W., Strong consistency under the Koziol-Green model. *Statistics & Probability Letters* **14** (1992a) 313-320.

Stute, W., Strong consistency of the MLE under random censorship. *Metrika* **39** (1992b) 257-267.

Stute, W., The bias of Kaplan-Meier integrals. *Scand. J. Statist.* **21** (1994) 475-484.

Stute, W., The central limit theorem under random censorship. *Ann. Statist.* **23** (1995) 422-439.

Stute, W. and J.-L. Wang, The strong law under random censorship. *Ann. Statist.* **21** (1993) 1591-1607.

Sun, L. and L. Zhu, A semiparametric model for truncated and censored data. *Statist. Probab. Letters* **48** (2000) 217-227.