A SIMPLE IMPROVEMENT OF THE KAPLAN-MEIER ESTIMATOR

Agnieszka Rossa

Dept. of Stat. Methods, University of Lódź, Poland
Rewolucji 1905, 41, Lódź
e-mail: agrossa@krysia.uni.lodz.pl

and

Ryszard Zieliński

Inst. Math. Polish Acad. Sc.
P.O.Box 137 Warszawa, Poland
e-mail: rziel@impan.gov.pl

ABSTRACT

Though widely used, the celebrated Kaplan-Meier estimator suffers from a disadvantage: it may happen, and in small and moderate samples it often does, that even if the difference between two consecutive times t_1 and t_2 ($t_1 < t_2$) is considerably large, for the values of the Kaplan-Meier estimators $KM(t_1)$ and $KM(t_2)$ at these times we may have $KM(t_1) = KM(t_2)$. Although that is a general problem in estimating a smooth and monotone distribution function from small or moderate samples, in the context of estimating survival probabilities the disadvantage is particularly annoying. In the paper we discuss a local smoothing of the Kaplan-Meier estimator based on an approximation by the Weibull distribution function. It appears that Mean Square Error and Mean Absolute Deviation of the smoothed estimator is significantly smaller. Also Pitman Closeness Criterion advocates for the new version of the estimator.

AMS 2000 subject classification: Primary 62N02 secondary 62G05

Key words and phrases: Kaplan-Meier estimator, Weibull distribution, survival probability

INTRODUCTION

Let $F(x), x \geq 0$, be the cumulative distribution function (CDF) of time to failure X of an item and let $G(y), y \geq 0$, be the CDF of random time to censoring Y of that item. Let $T = min(X, Y)$, let $I(A)$ denote the indicator function of the set A, and let $\delta = I(X \leq Y)$. Given $t > 0$, the problem is to estimate survival distribution function (SDF) $\bar{F}(t) = 1 - F(t)$ from the "incomplete" ordered sample

$$(1) \qquad (T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n), \quad T_1 \leq T_2 \leq \ldots \leq T_n$$

The Kaplan-Meier (1958) estimator (KM), also called the product limit estimator, is defined as

$$(2) \qquad KM(t) = \begin{cases} \prod_{i=1}^{n} \left(1 - \dfrac{\delta_i}{n-i+1}\right)^{I(T_i \leq t)}, & \text{for } t \leq T_n \\ \begin{cases} 0, & \text{if } \delta_n = 1 \\ \text{undefined}, & \text{if } \delta_n = 0 \end{cases}, & \text{for } t > T_n \end{cases}$$

In the case of ties among the T_i we adopt the usual convention that failures ($\delta_i = 1$) precede censorings ($\delta_i = 0$). By the definition, KM estimator is right-continuous.

Efron (1967) modified the estimator defining $KM(t) = 0$ where originally it was not defined; Gill (1980) proposed another modification with $KM(t) = KM(T_n)$, where the Kaplan–Meier estimator in its original version was not defined (i.e. for $t > T_n$ when $\delta_n = 0$).

To get some intuition concerning these versions and to illustrate our approach we shall refer to the well know example from Freireich *at al.* (1963) - see also Peterson (1983) or Marubini and Valsecchi (1995). The "survival times" of 21 clinical patients were

$$(3) \quad 6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*, 34^*, 35^*$$

where * denotes a censored observation. Kaplan-Meier estimator for that data is presented in Fig. 1.

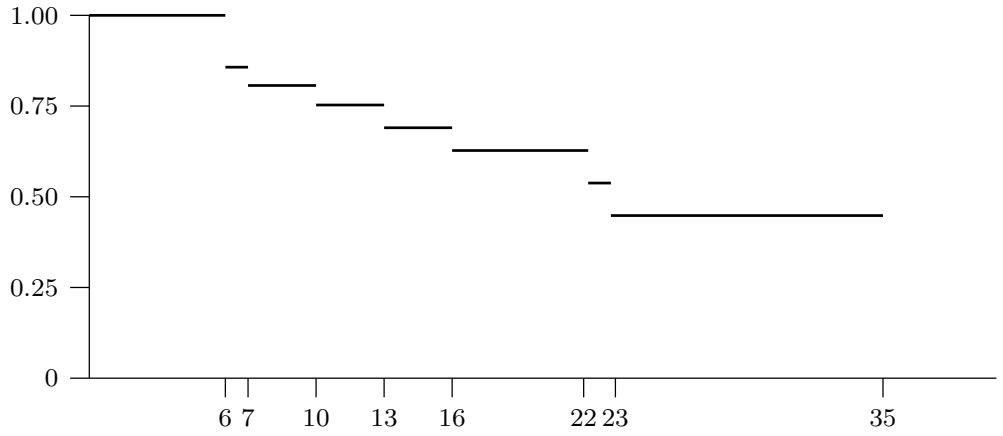

Fig.1. Kaplan-Meier estimator for data (3)

A disadvantage of those estimators is that in small and moderate samples it may happen, and it often does, that even if the difference between two different times t_1 and t_2 $(t_1 < t_2)$ is considerably large, for the values of the Kaplan-Meier estimators $KM(t_1)$ and $KM(t_2)$ at these times we may have $KM(t_1) = KM(t_2)$. For example, for the above data we have $KM(17) = KM(20) = 0.627$ and

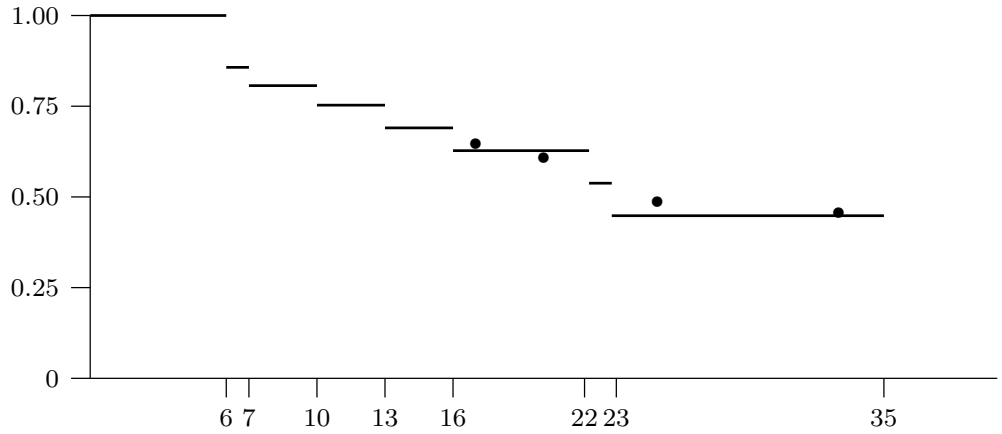

Fig.2. Estimated values (dots) of estimator S_2 for data (3)

$KM(25) = KM(33) = 0.448$. It is really very difficult for a statistician to explain to a practitioner why the probability to survive at least $t = 25$ is equal to the probability of surviving at least $t = 33$! The estimator we propose, denoted by S_2,

gives us $S_2(17) = 0.6451$, $S_2(20) = 0.6065$, $S_2(25) = 0.4842$, and $S_2(33) = 0.4545$ (see Fig. 2) which obviously sounds more reasonably.

Another disadvantage of the Efron and Gill estimators is that they estimate the survival probability beyond what one can reasonably conclude from the sample. It is obvious that Efron guessing will be preferable for short-tailed distributions ("a pessimistic prophet") and Gill for the fat-tailed distributions ("an optimistic prophet") but to reasonable choose between them one should restrict in a way the original nonparametric model. For that reason we confine ourselves to the original Kaplan-Meier version (2).

LOCAL WEIBULL SMOOTHING

Kaplan-Meier estimator is adequate for the nonparametric statistical model in which the only assumptions concerning possible distributions of life time are their continuity and strict monotonicity. There are some well known representatives of that family of distributions:

– exponential $E(\lambda)$ with probability density function PDF $\propto \exp\{-\lambda t\}$
– Weibull $W(\lambda, \alpha)$ with SDF $W(t; \lambda, \alpha) = \exp\{-\lambda t^\alpha\}$
– gamma $\Gamma(\lambda, \alpha)$ with PDF $\propto t^{\alpha-1} \exp\{-\lambda t\}$)
– generalized gamma $\Gamma_g(\lambda, \alpha, k)$ with PDF $\propto t^{\alpha k-1} \exp\{-\lambda t^\alpha\}$
– lognormal $logN(\mu, \sigma)$
– Gompertz $Gom(\lambda, \alpha)$ with SDF of the form $exp\{\lambda(1 - exp(\alpha t))\}$
– Pareto $Par(\lambda, \alpha)$ with SDF equal to $(1 + \lambda t)^{-\alpha}$
– log-logistic $logL(\lambda, \alpha)$ with SDF $1/(1 + \lambda t^\alpha)$
– exponential-power $EP(\lambda, \alpha)$ with PDF $\propto exp\{-\lambda t^\alpha\}$

to mention the most popular among them (e.g. Kalbfleisch and Prentice 1980, Klein et al. 1990). Here $\propto$ means as usually "proportional to".

It is obvious that on a sufficiently short interval on the real half-line each of them may be considered as a reasonably good approximation of any distribution function F from the nonparametric family under consideration. To approximate the tail $\bar{F} = 1 - F$ we have chosen the Weibull tail $W(t; \lambda, \alpha) = exp\{-\lambda t^\alpha\}$ mainly

because that gives us a simple algorithm of calculating the estimator: it is enough to perform a suitable logarithmic transformations of data and apply the standard estimating procedure for A and B in the simple regression model $y = Ax + B$.

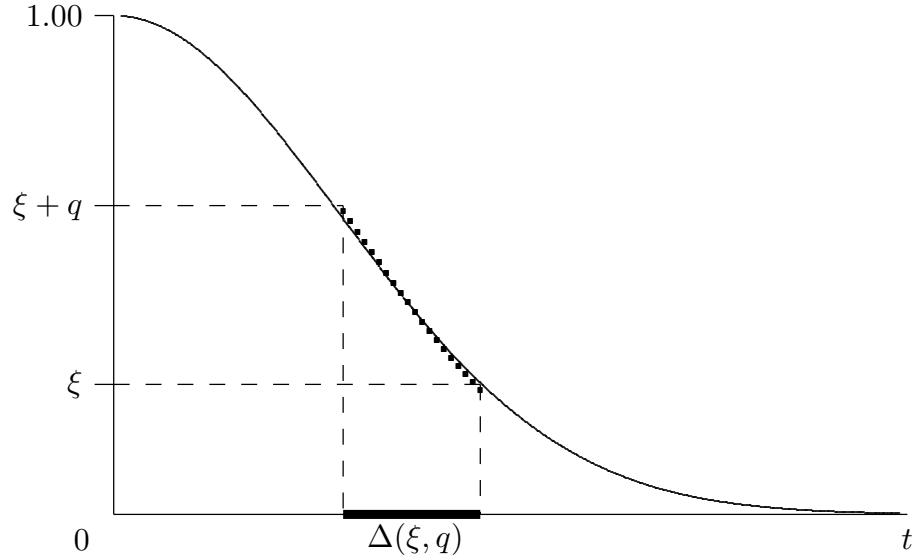

Fig.3. SDF (solid line) and its local Weibull approximation on $\Delta(\xi, q)$ (dotted line)

On the other hand, the Weibull family $\left\{exp\{-\lambda t^{\alpha}\}, \lambda > 0, \alpha > 0\right\}$ of tails appears to be sufficiently flexible to locally fit all typical survival distributions. To see that take any CDF F and its SDF $\bar{F} = 1 - F$, fix $q \in (0, 1)$, and consider the family of all intervals $[\xi, \xi + q)$, $0 \le \xi \le 1 - q$. For a fixed $\xi \in [0, 1 - q]$ denote by $\Delta(\xi, q)$ the interval $[\bar{F}^{-1}(\xi + q), \bar{F}^{-1}(\xi)]$ on the real halfline (Fig.3). Let

$$\mu_F(\xi, q) = \min_{\lambda,\alpha} \max_{t \in \Delta(\xi,q)} |\bar{F}(t) - W(t; \lambda, \alpha)|$$

The quantity $\mu_F(\xi, q)$ tells us how good is the best Weibull approximation of the SDF $\bar{F}$ on the interval $\Delta(\xi, q)$. It is obvious that for every $\varepsilon > 0$ one can find a sufficiently small q such that

$$\mu_F(q) := \max_{0 \le \xi \le 1-q} \mu_F(\xi, q) < \varepsilon$$

Observe that $\mu_F(1)$ is the error (in the "sup–norm") of approximation of F by the best Weibull CDF on the whole positive halfline. Numerical values of $\mu_F(q)$ for some q and for a number of representatives F are given in Tab.1.

Tab.1

F	$\mu_F(q)$					
	$q=1$	$q=0.10$	$q=0.20$	$q=0.25$	$q=0.30$	$q=0.50$
$\Gamma(1,2)$	0.0115	0.0006	0.0018	0.0021	0.0026	0.0097
$logN(0,1)$	0.0381	0.0018	0.0043	0.0058	0.0075	0.0162
$Gom(1,1)$	0.0219	0.0011	0.0027	0.0036	0.0046	0.0102
$Par(1,2)$	0.0222	0.0021	0.0044	0.0057	0.0070	0.0127
$logL(1,1)$	0.0389	0.0031	0.0066	0.0087	0.0110	0.0218
$EP(1,2)$	0.0151	0.0005	0.0013	0.0019	0.0062	0.0063

Now the idea of an estimator is: choose q which gives us a satisfactory level of the error of approximation of any F from a family under consideration and, to estimate survival probability at a point t, smooth the Kaplan–Meier estimator in a vinicity of t, which contains $100q$ per cent of sample points.

THE ESTIMATOR

Let $N-1$ be the number of distinct elements of the sample (1) in which $\delta_i=1, i<n$, and let $i_1, i_2, \ldots, i_{N-1}$ be indexes of those elements. Let $T'_0 = 0$ and define $T'_j = T_{i_j}$ and $T'_N = T_n$. Then $KM(T'_0) = 1$ and $KM(t), t < T_n$, has jumps at points $T'_j, j = 1, 2, \ldots, N-1$, and only at these points. If $\delta_n = 1$, then also $t = T_n$ is a point of a jump of KM. We shall write Kaplan-Meier estimator in the form of the sequence of pairs $(T'_j, KM'_j), j = 1, 2, \ldots, N)$, where

$$KM'_j = \begin{cases} \dfrac{KM(T'_{j-1}) + KM(T'_j)}{2}, & \text{if } j = 1, 2, \ldots, N-1 \\ \begin{cases} KM(T_n)/2 & \text{for } \delta_n = 1 \\ KM(T_n) & \text{for } \delta_n = 0 \end{cases}, & \text{if } j = N \end{cases} \tag{4}$$

For data (3) we have $N = 8$ and $(6, 0.929)$, $(7, 0.832)$, $(10, 0.780)$, $(13, 0.722)$, $(16, 0.659)$, $(22, 0.583)$, $(23, 0.493)$, and $(35, 0.448)$.

Suppose we want to estimate survival probability $P\{X > t\}$ at a point t. If $t > T_n$ and $\delta_n = 0$, our estimator, like the original Kaplan-Meier estimator (2), is not defined. Otherwise we conctruct our estimator as follows.

The smallest number of points to fit a two-parameter Weibull curve $W(t; \lambda, \alpha) = \exp\{-\lambda t^\alpha\}$ equals 2. Let us begin with construction of an estimator based on two consecutive observations T'_j.

It follows from Tab.1 that, for example, the maximal error of the local approximation of $\log N(0, 1)$ distribution on each interval that contains $100q = 25$ per cent of the population equals 0.0058 and that for no representative distribution the error exceeds 0.01. If we accept that level of the error of approximation as satisfactory, we may agree to base our estimation on two consecutive observations T'_{k-1}, T'_k such that $T'_{k-1} \le t \le T'_k$, whenever two observations make no more than 25 per cent of the sample, i.e. whenever $N \ge 8$, and estimate the survival probability at t by the value $W(t; \lambda, \alpha)$ with λ and α chosen in such a way, that $W(T'_{k-1}; \lambda, \alpha) = KM'_{k-1}$ and $W(T'_k; \lambda, \alpha) = KM'_k$,

Generally: if, to control the error of local approximation we have decided to choose q, we may believe that whenever $N \ge 2/q$, the error of the estimator based on two consecutive observations would not exceed $\max_F \mu_F(q)$. (The total error of the estimator, that includes the error of estimation of an unknown SDF by a Weibull tail, a systematic error caused by censoring, and a random error of the sample, is of course greater.) This leads us to the following estimator $S_2(t)$:

$$S_2(t) = \exp\{-\exp(Y)\} \tag{5}$$

where

$$Y = \begin{cases} Y_1 + \dfrac{Y_2 - Y_1}{X_2 - X_1}(X - X_1), & \text{if } T'_0 \le t \le T'_1 \\ Y_{k-1} + \dfrac{Y_k - Y_{k-1}}{X_k - X_{k-1}}(X - X_{k-1}), & \text{if } T'_{k-1} \le t \le T'_k \le T'_N \\ Y_{N-1} + \dfrac{Y_N - Y_{N-1}}{X_N - X_{N-1}}(X - X_{N-1}), & \text{if } t > T'_N \text{ and } \delta_N = 1 \\ \text{undefined} & \text{if } t > T'_N \text{ and } \delta_N = 0 \end{cases}$$

and

$$X_j = \log(T'_j), \quad Y_j = \log(-\log(KM'_j)), \quad X = \log t$$

The estimator is presented in Fig.4.

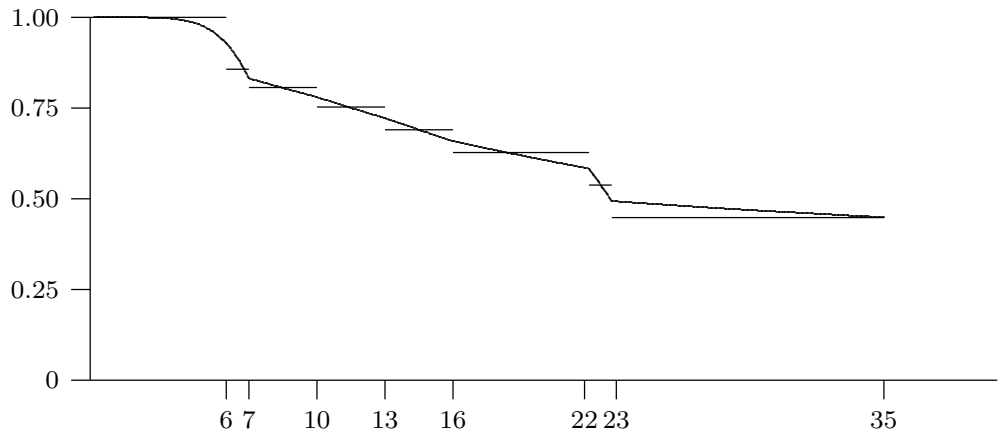

Fig.4. Kaplan-Meier and S_2 estimators for data (3)

One may expect that the estimator based on $m > 2$ neighbours of t would be more smooth and more exact. It is really so but a flaw of this estimator makes a difficulty in its practical applications. Let us begin with a construction of such estimator.

First, choose $\varepsilon > 0$ as a satisfactory level of the error of local approximation of a survival probability by a Weibull tail and find q (see previous Section). Then, by arguments as above, take $m = [qN]$. We should obtain $m \geq 2$. If that is not the case, we are not able to estimate survival probability at t within the prescribed accuracy ε of local approximation of SDF by a Weibull tail.

If $m = 2k$ is even, define

$$w = \begin{cases} 1, & \text{if } t < T'_k \\ j - k + 1, & \text{if } T'_{j-k+1} < \ldots < T'_j \leq t < T'_{j+1} < \ldots, < T_{j+k} \\ N - m + 1, & \text{if } T'_{N-k+1} \leq t \end{cases}$$

If $m = 2k + 1$ is odd, find T'_{j^*} such that $|T'_{j^*} - t| \leq |T'_j - t|$, $j = 1, 2, \ldots, N$, and define

$$w = \begin{cases} 1, & \text{if } j^* \leq k+1 \\ j^* - k, & \text{if } k+1 < j^* \leq N-k \\ N - m + 1, & \text{if } N-k < j^* \end{cases}$$

Take $T'_w, T'_{w+1}, \ldots, T'_{w+m-1}$ as neighbours of the point t. Then fit a Weibull tail $\exp\{-\lambda t^\alpha\}$ to them. To this end "linearize the tail" by introducing auxiliary variables

$$x_j = \log(KM'_j), \quad y_j = \log(-\log T'_j), \quad j = w, w+1, \ldots, w+m-1$$

and estimating regression coefficients Λ and α in

$$y = \Lambda + \alpha x$$

where $\Lambda = \log \lambda$. Finally, if $(\hat{\Lambda}, \hat{\alpha})$ are estimators of those coefficients and $\hat{\lambda} = \exp\{\hat{\Lambda}\}$, estimate survival probability $P\{X > t\}$ by

$$S_{[qN]}(t) = \exp\{-\hat{\lambda} t^{\hat{\alpha}}\} \tag{6}$$

As it was expected, the estimator based na $m > 2$ points is more accurate (see next Section), however a disadvantage of the smoothed estimator based on $m > 2$ points consists in that it may happen, and sometime it does (see Fig.5), that $sKM(t_1) < sKM(t_2)$ though $t_1 < t_2$. It may happen if t_1 has a value close to the upper bound of the interval of those t, which are estimated by smoothing the points at $T'_w, T'_{w+1}, \ldots$ and t_2 is close to the lower bound of the interval of those t, which are estimated by smoothing the points at $T'_{w+1}, T'_{w+2}, \ldots$. In that case a kind of adjustment of the estimator at two adjacent points is needed but as yet we do not how to approach the problem.

Like the original Kaplan-Meier estimator KM, the smoothed estimators (5) and (6) are difficult for theoretical analysis. It is obvious that for large n and in consequence for large N, the estimators will behave like KM, however in an asymptotic setup one can hardly expect new interesting results.

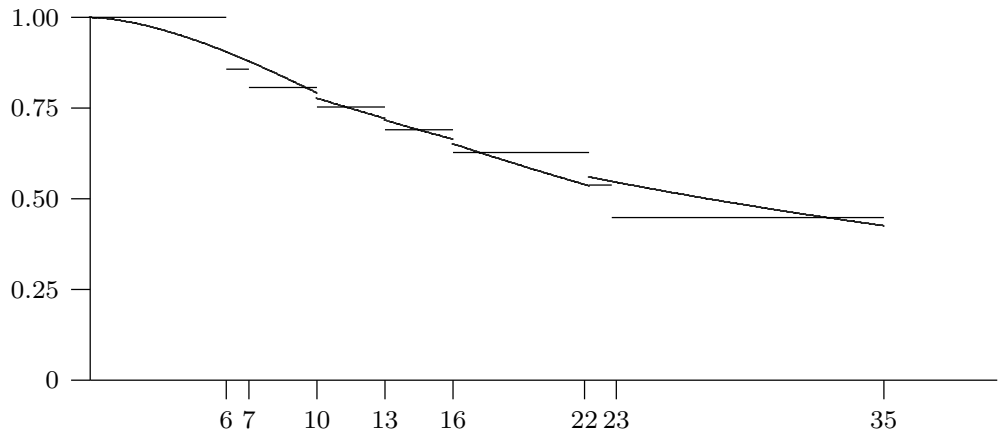

Fig.5. Kaplan-Meier and $S_{[0.25N]}$ estimators for data (3)

In small and moderate samples the smoothed estimator may considerably differ from the original one, in such situations however general theoretical conclusions seem to be impossible. Simulation studies (next Section) demonstrates that the proposed smoothing really improves estimation.

A SIMULATION STUDY

We have chosen $q = 0.25$. To compare the estimators for a fixed time-to-failure distribution F and for a fixed censoring distribution G, we generated n independent observations $X_1, \ldots, X_n$ from F and n independent observations $Y_1, \ldots, Y_n$ from G. Next we calculated the ordered sample of the form (1) and the Kaplan-Meier estimator in the form (4). If $N < 8$, the sample was rejected. We continued simulation until we observed the prescribed number L of samples. In all simulations we assumed $L = 10,000$. For a given $p \in (0,1)$ we estimated survival probability by the Kaplan-Meier estimator $KM(t)$ and by estimators $S_2(t)$ and $S_{[0.25N]}(t)$ for $t = \bar{F}^{-1}(p)$. For each estimator, we calculated Mean Squar Error (MSE) and Mean Absolute Deviation (MAD). As it was expected, for large sample sizes the estimators do not differ substantially and in small samples estimators S_2 and $S_{[0.25N]}$ prevail. **In an exhaustive simulation study we found that** $0.47 \leq MSE(S_{[0.25N]})/MSE(KM) \leq 0.89$ **with the lower bound** 0.47 **observed for Gompertz** $Gom(1,1)$ **distribution as** F **and exponential distribution** $E(3)$ **as** G**, both for samples** $n = 10$**, which means that mean**

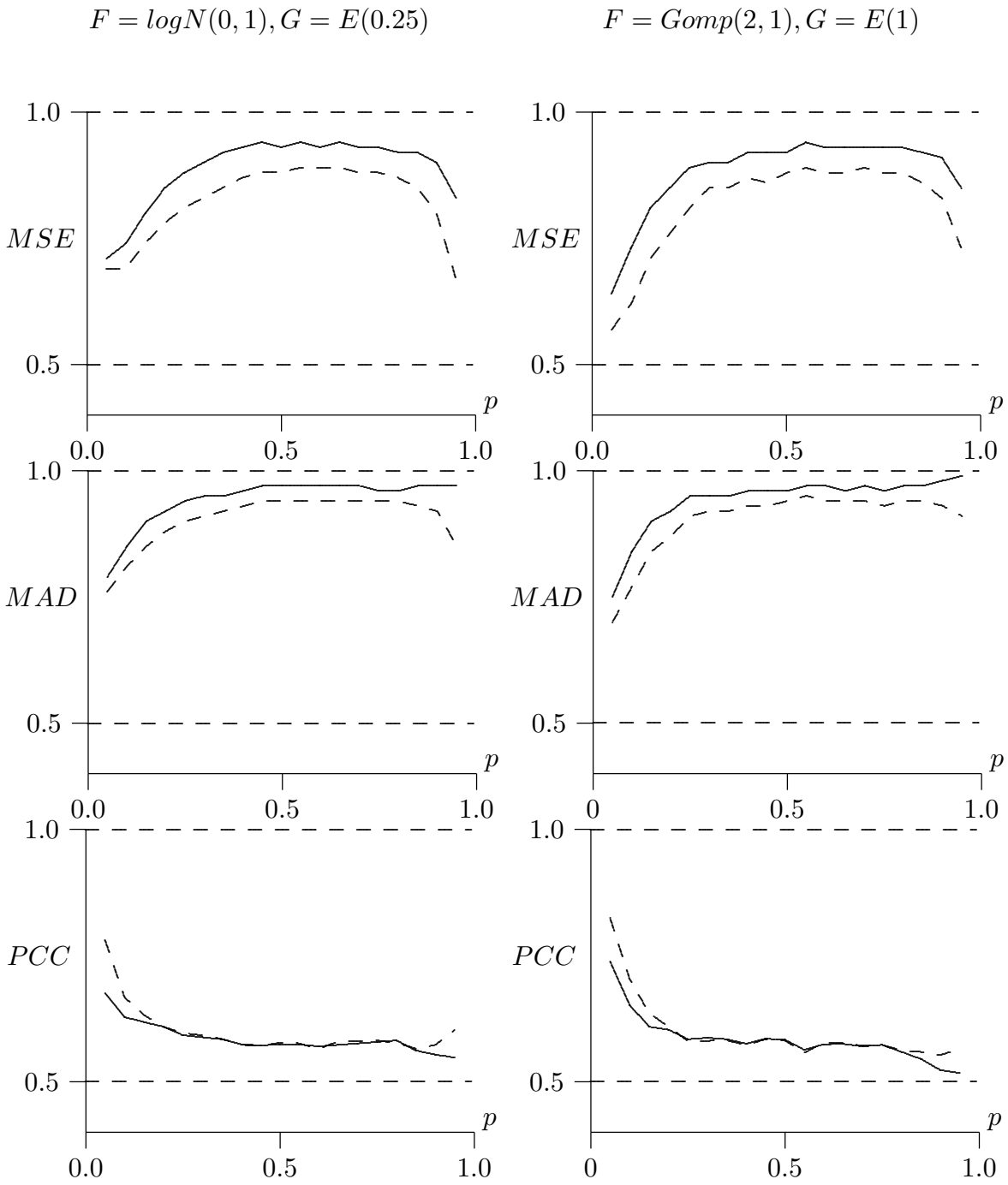

Fig.6. Simulated ratios of MSE and MAD of smoothed estimators S_2 (solid lines) and $S_{[0.25N]}$ (dashed lines) to those of the Kaplan-Meier estimator. PCC_2 - solid lines, $PCC_{[0.25N]}$ - dashed lines.

square error was reduced by 11 to 53 per cent. Similar reduction for MAD was between 9 (for Pareto $Par(2,2)$ and exponential $E(1)$ distributions) and 39 (for $Gom(1,1)$ and $E(1)$) per cent. More numerical results have been presented in a technical report (Rossa and Zieliński 1999).

Typical results that we have obtained for $p \in (0,1)$ are presented in Fig. 6, where the ratios of mean square errors $MSE(S_2)/MSE(KM)$ (solid lines) and $MSE(S_{[0.25N]})/MSE(KM)$ (dashed lines) are exhibited for two pairs of the time-to-failure distribution F and censoring distribution G.

Let $PCC_2(p,F,G)$ denote the Pitman Closeness Criterion (see Keating et al. 1993) for estimators S_2 and KM at the point $t = \bar{F}^{-1}(p)$ if $\bar{F}$ is the survival distribution and G is censoring distribution:

$$PCC_2(p,F,G) = P_{F,G}\{|S_2(t) - p)| \leq |KM(t) - p)|\}$$

If $PCC_2(p,F,G) > 0.5$ then S_2 prevails in the sense that the absolute error of this estimator is smaller than that of KM more often than it is larger. Similar notation $PCC_{[qN]}$ we adopt for estimator $S_{[qN]}$.

Fig. 6 exhibits typical behaviour of PCC for the whole range of $p \in (0,1)$.

ADDITIONAL COMMENTS

1. Our simulations suggest that also the bias of our estimators is smaller than that of KM but we were not able to find any regularity in that. Whatever however the bias and the variance, with respect to MSE and MAD the smoothed estimators are better than the original KM.

2. Sufficiently far to the right, the original Kaplan-Meier estimator in all simulations gives the values zero. It follows that "practically", for large t, it's variance is equal to zero. That of course is not the case for the smoothed estimators.

3. In our simulations we were also interested in MSE and other characteristics of the estimators under consideration if there is no censoring. For every $p \in (0,1)$ the Kaplan-Meier estimator is then unbiased and is variance at the point $\bar{F}^{-1}(p)$ is equal to $p(1-p)/n$. It is interesting to observe (Fig. 7) that sometimes censoring may improve MSE. On a paradox of this kind see Csörgö et al. (1998).

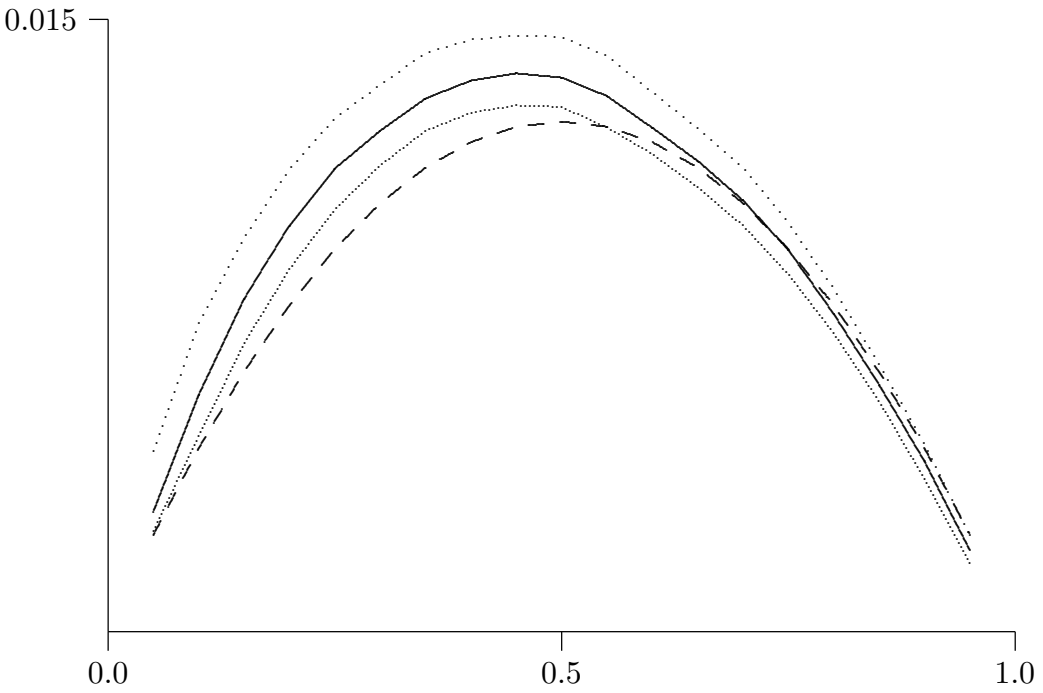

Fig.7. Simulated MSE of the Kaplan-Meier estimator (dotted line), the proposed smoothed estimator S_2 (solid line), $S_{[0.25N]}$ (dense dotted line) and Kaplan-Meier estimator without censoring (dashed line). $F = Gom(2,1)$, $G = E(0.75)$, sample size $n = 20$.

ACKNOWLEDGEMENTS

The research of the second author has been supported by grant KBN 2 P03A 033 17.

REFERENCES

Csörgö, S. and Faraway, J.J. (1998). The paradoxical nature of the proportional hazards model of random censorship. *Statistics* 31, 67-78

Efron, B. (1967). The two-sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 4, 831–852

Freireich, E.O. *et al.* (1963). The effect of 6-mercaptopurine on the duration of steroid-induced remission in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood* 21, 699–716

Gill, R.D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tract No. 124, Amsterdam: Mathematisch Centrum.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data*. Wiley

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Keating, J.P., Mason, R.L, and Sen, P.K. (1993). Pitman's Measure of Closeness: A comparison of Statistical Estimators SIAM Philadelphia

Klein, J.P., Lee, S.C. and Moeschberger, M.L. (1990). A partially parametric estimator of survival in the presence of randomly censored data. *Biometrics*, 46, 795–811.

Marubini, E. and Valsecchi, M.G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*, Wiley

Peterson, A.V. (1983). Kaplan-Meier estimator. In *Encyclopedia of Statistical Sciences*, S. Kotz, N.L.Johnson, and C.B.Read, eds., Vol.4, Wiley

Rossa, A. and Zieliński, R. (1999). Locally Weibull-Smoothed Kaplan–Meier Estimator, *Institute of Mathematics Polish Academy of Sciences, Preprint 599*