SORT 32 (1) January-June 2008, 67-76

Empirical comparison between the Nelson-Aalen Estimator and the Naive Local Constant Estimator

Ana M. Pérez-Marín[*]

University of Barcelona

Abstract

The Nelson-Aalen estimator is widely used in biostatistics as a non-parametric estimator of the cumulative hazard function based on a right censored sample. A number of alternative estimators can be mentioned, namely, the *naive local constant estimator* (Guillén, Nielsen and Pérez-Marín, 2007) which provides improved bias versus variance properties compared to the traditional Nelson-Aalen estimator. Nevertheless, an empirical comparison of these two estimators has never been carried out. In this paper the efficiency performance of these two estimators when applied to real survival data are compared. Our results suggest that the efficiency improvement introduced by the naive local constant estimator is highly remarkable for all distribution quantiles, especially for low quantiles.

MSC: 62N01, 62N02

Keywords: Cumulative hazard function, efficiency, Nelson-Aalen estimator, survival analysis.

1 Introduction

The Nelson-Aalen estimator was first introduced by Nelson (1972) in a reliability context and later on rediscovered by Aalen (1978) who derived the estimator using modern counting process techniques (Klein and Moeschberger, 1993). This estimator has a number of nice properties (Andersen, Borgan, Gill & Keiding, 1993) and better small-sample-size performance than other standard methods. The properties and efficiency performance of the Nelson-Aalen estimator are the topic of a number of papers (see Peña & Rohatgi, 1993, among may others). It is very popular in biostatistics

[*] Department of Econometrics, RFA-IREA. University of Barcelona. Av. Diagonal 690. 08034 Barcelona. amperez@ub.edu.
Received: February 2007
Accepted: January 2008

and it is basically used in two different ways when analysing survival data (Klein and Moeschberger, 1997) – firstly, it provides useful information in order to select between parametric models for the time-to-event variable, and secondly, it provides crude estimates of the hazard rate which can be subsequently smoothed.

Guillén, Nielsen and Pérez-Marín (2007) proved that the efficiency of the Nelson-Aalen estimator can be considerably improved by using more information in the estimation process than that used by the traditional Nelson-Aalen estimator uses. When the Nelson-Aalen estimator is estimated at a point t (indicating the time-to-event) it only uses information on the interval $[0, t]$. Guillén, Nielsen and Pérez-Marín (2007) proved that this estimator can be improved by using some information just to the right of t, in $[t, t + b]$, where the bandwidth parameter b is small and depends on t. In this way, some bias is introduced but the variance is reduced at the same time. The goal of the authors is to formulate the new estimator and to obtain the optimal b resulting in an overall efficiency gain compared to the Nelson-Aalen estimator. The efficiency gain is measured by the absolute efficiency gain of the naive local constant estimator with respect to the Nelson-Aalen estimator, divided by the efficiency of the Nelson-Aalen estimator, i.e. the relative efficiency gain of the new estimator with respect to the standard Nelson-Aalen estimator.

2 Some remarks on the naive local constant estimator

Guillén, Nielsen and Pérez-Marín (2007) adapted the model formulation of Andersen, Borgan, Gill & Keiding (1993, p. 176) with an infinite terminal point $\tau = \infty$ not included in the considered interval, $[0, \infty[$. They considered a measurable space (Ω, F), equipped with a filtration $(F_t, t \in [0, \infty[)$ which satisfies the usual conditions except for possible completeness, see Andersen, Borgan, Gill & Keiding (1993, p. 60), for each member of a family P of probability measures. The authors also considered the multivariate counting process $N = \{N_1(t), \ldots, N_n(t)\}$, $t \in [0, \infty[$ that satisfies Aalen's multiplicative intensity model, i.e., its (P, F_t)–intensity process is $\lambda_i(t) = \alpha(t)Y_i(t)$, where α is the hazard function and Y_i is an observable predictable process taking values in $\{0, 1\}$, indicating, by the value 1, when the i th individual is under risk. If $Y = \sum_{i=1}^{n} Y_i$ is the aggregated exposure, the sum of individual processes indicating that the unit is under risk, the Nelson-Aalen estimator for the cumulative hazard $\Lambda(t)$ equals

$$\widehat{\Lambda}_{NA}(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{Y(s)} dN_i(s).$$

The authors provide the general formulation of the naive local constant estimator, given by $\widehat{\Lambda}_{NLC}(t) = \int_0^\infty w(s,t)d\widehat{\Lambda}_{NA}(s)$, for all t, where $w(s,t)$ is some weight function. When considering a "naive" weight function, i.e. uniform weighting in the neighbourhood of t ($w(s,t) = I_{\{s\leq t-b\}} + \frac{1}{\gamma_{t,b}} I_{\{s\in(t-b,t+b)\}}$, where I is the indicator function) the estimator can be expressed as follows (Guillén, Nielsen & Pérez-Marín, 2007):

$$\widehat{\Lambda}_{NLC}(t) = \widehat{\Lambda}_{NA}\left\{\max(t-b,0)\right\} + \frac{1}{\gamma_{t,b}}\left[\widehat{\Lambda}_{NA}(t+b) - \widehat{\Lambda}_{NA}\left\{\max(t-b,0)\right\}\right]$$

where

$$\gamma_{t,b} = \frac{t+b-\max(t-b,0)}{t-\max(t-b,0)}.$$

According to the authors, the term "naive" was taken from Silverman (1986), who used it for a kernel density estimator, where the kernel equalled the uniform distribution.

In this context, Guillén, Nielsen and Pérez-Marín (2007) found that the optimal bandwidth parameter (the one maximizing the relative efficiency gain with respect to the Nelson-Aalen estimator) equals $b_{opt} = \{\alpha(t)/\left[2Y(t)\alpha'(t)^2\right]\}^{\frac{1}{3}}$ for $t \geq b$ and $b_{opt} = t$ in the boundary region, when $t < b$. The corresponding relative efficiency gain $\varepsilon(t)$ equals

$$\varepsilon(t) = \begin{cases} \dfrac{3}{8}\dfrac{\alpha(t)}{\displaystyle\int_0^t \alpha(s)ds} b_{opt} & \text{if } b_{opt} \leq t \\[2ex] \dfrac{\dfrac{t}{2}\left\{\alpha(t) - \dfrac{Y(t)t^3}{2}\alpha'(t)^2\right\}}{\displaystyle\int_0^t \alpha(s)ds} & \text{if } b_{opt} > t. \end{cases} \tag{1}$$

The main hypothesis behind the definition of this new estimator is that the hazard rate is locally constant around t, the point where the cumulative hazard is estimated. Additionally, it is assumed that the hazard function α does not depend on i, is twice continuously differentiable, $\int_0^t \alpha(s)ds < \infty$, $\int_0^t \alpha(s)ds \neq 0$ and $\alpha'(t) \neq 0$ for all $t \in [0,\infty[$.

The authors provide in their paper some examples of the efficiency comparison between these estimators by assuming some well-known distributions for the time-to-event variable. Nevertheless, they do not provide any empirical application where the performance of both estimators would be illustrated by using real survival data. An empirical study is necessary in this case to address some practical aspects of the implementation of this new estimator in survival studies, such as the estimation of the optimal bandwidth parameter based on real survival data and the efficiency improvement of the new estimator with respect to the Nelson-Aalen estimator.

3 The survival dataset

In this section we present the survival data set used in this paper in order to analyse the performance of the Nelson-Aalen estimator and the naive local constant estimator. This data set has been previously used (Andersen, Borgan, Gill & Keiding, 1993, example IV.1.2) to illustrate how to use the Nelson-Aalen on the basis of real survival data.

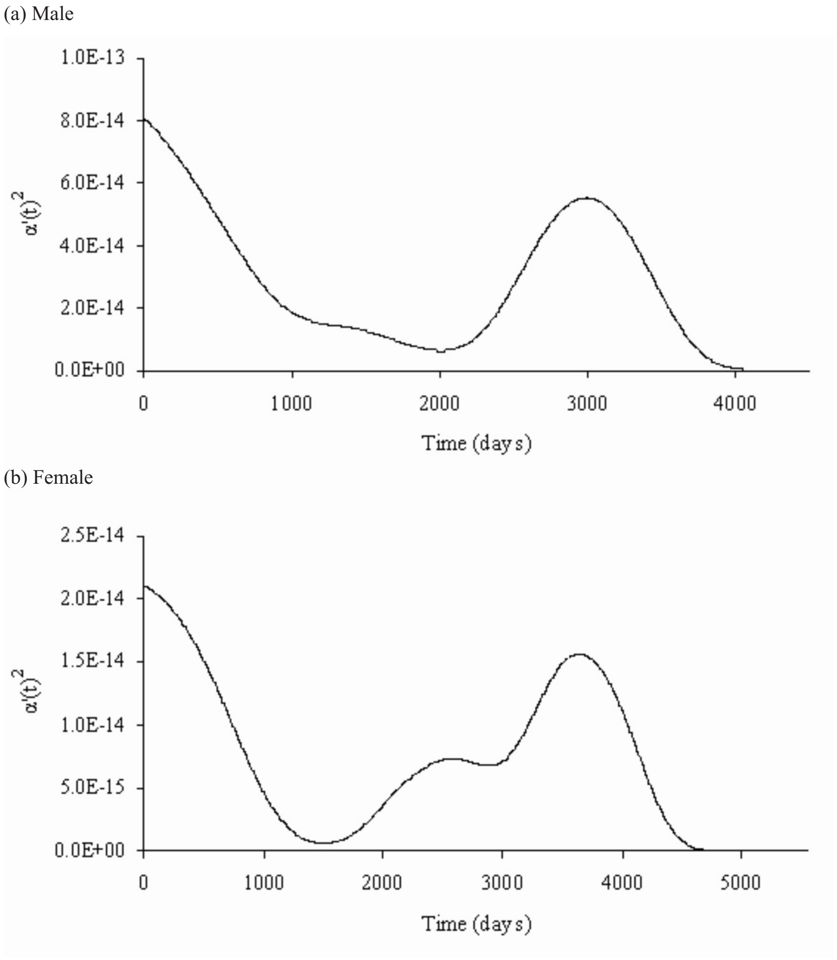
In the period from 1962 to 1977, 79 male and 126 female patients with malignant melanoma, cancer of the skin, had radical operations performed at the Department of Plastic Surgery, University Hospital of Odense, Denmark. The tumour was completely removed together with the skin within a distance of about 2.5 cm around it. All patients were monitored until the end of 1977 and it was noted if and when any of the patients died, as well as the cause of death. Of the 79 male patients, 29 were observed to die from the disease, and of the 126 female patients, 28 were observed to die from the disease, while 14 died from other causes. The rest of them were alive at the end of 1977. The objective of this historically prospective clinical study was to assess the effect of risk factors on survival. The most important time variable is time since operation. Other factors were screened such as gender, age at operation and several variables related to the characteristics of the tumour.

Andersen, Borgan, Gill & Keiding (1993, example IV.1.2) present Nelson-Aalen estimates for these male and female patients where the survival time is measured since the time of operation. We will now compare their results with those corresponding to the naive local constant estimator.

4 Estimating the optimal bandwidth

In this section we estimate the optimal bandwidth parameter by following the procedures provided by Guillén, Nielsen and Pérez-Marín (2007). In order to get the optimal bandwidth b, both $\alpha(t)$ and $\alpha'(t)^2$ should be estimated first. According to the authors, these estimators can be obtained by using the local linear estimator. Nielsen & Tanggaard (2001) introduced local linear hazard estimation based on "locally" fitting a line to the survival data via weighted least square kernel estimation. The slope of this line at each survival time let us to estimate $\alpha'(t)^2$.

Regarding the probability function $K_b(\cdot)$, we calculated the efficiency improvement for different well-known kernel functions and finally we used the biweight kernel $K_b(\cdot) = \frac{15}{16}\{1-(\cdot/b)^2\}^2$ where $b = 800$ for both male and female as it provides a substantial efficiency improvement. The same biweight kernel with the same b has been used to smooth $\alpha'(t)^2$ one more time (as explained in Guillén, Nielsen and Pérez-Marín, 2007). In this way, a more robustified estimator for $\alpha'(t)^2$ has been obtained, see Figure 1.


(a) Male

(b) Female

Figure 1: *Estimation of $\alpha'(t)^2$.*

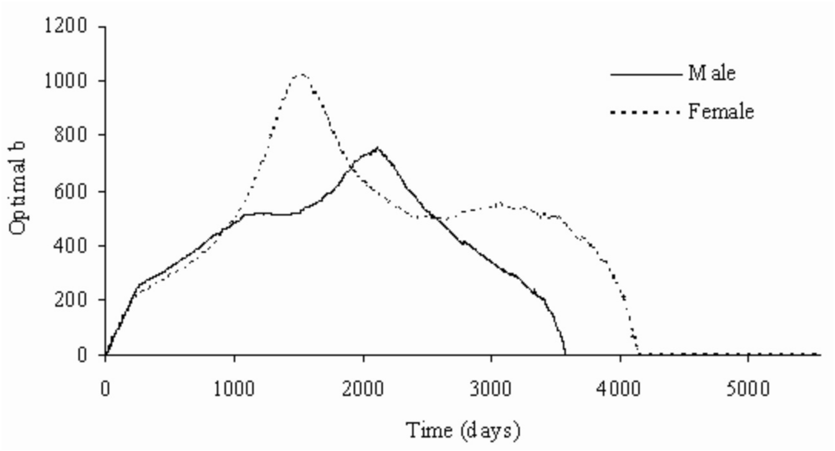


Figure 2: *Optimal b used in the naive local constant estimator for male and female.*

Once estimators of $\alpha(t)$ and $\alpha'(t)^2$ have been obtained the optimal b can easily be calculated. In Figure 2 optimal b as a function of t is shown both for male and female. Note that for small survival times the optimal bandwidth is exactly equal to this survival time (as proved by Guillén, Nielsen and Pérez-Marín, 2007). It is important to remark that for each t the optimal bandwidth b_{opt} determines the neighbourhood around t where the hazard is assumed to be constant. The largest value of this bandwidth corresponds to the time point $t = 2000$ days and $t = 1500$ days approximately for male and female respectively.

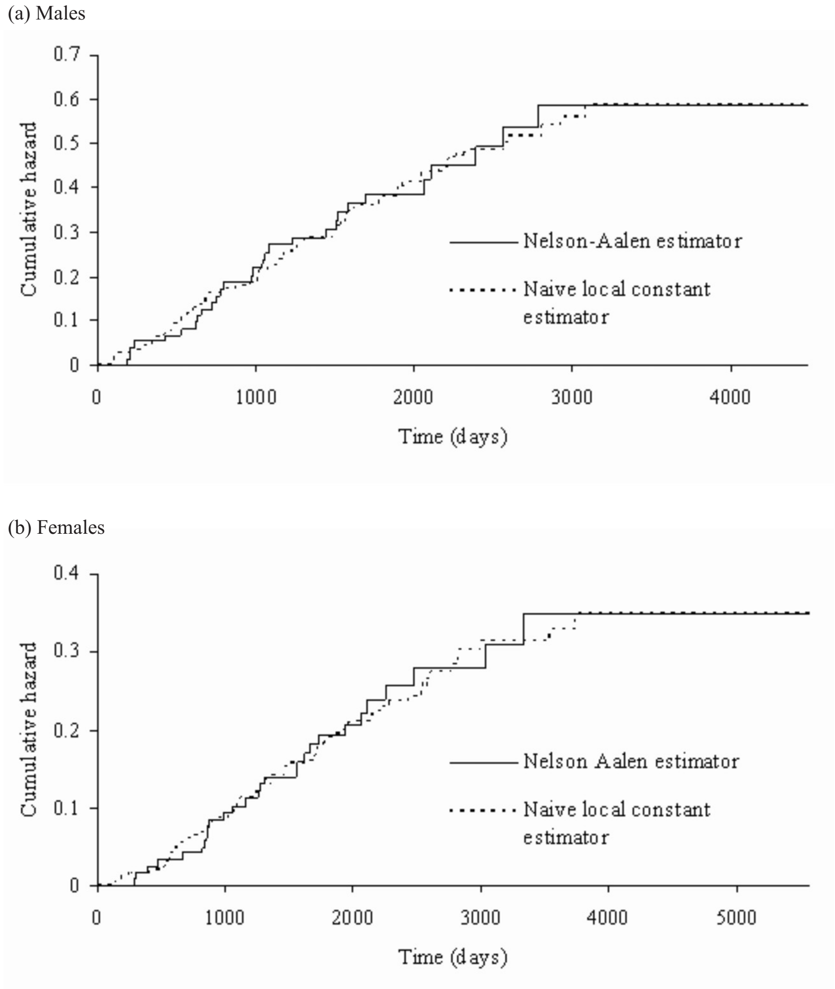
It is important to remark that, in order to get the estimation of the bandwidth parameter b_{opt}, it is necessary to estimate both $\alpha(t)$ and $\alpha'(t)^2$. This could be viewed as a practical limitation. Nevertheless, it is necessary to notice that the worst that could happen in the case where it is not possible to have suitable estimators of these two functions would be to have neither the optimal estimation of b nor the optimal efficiency improvement with respect to the Nelson-Aalen estimator, but some approximation. Therefore, we would lose part of the efficiency improvement, but not all of that, as (1) is always positive.

5 The cumulative hazard function estimates

In Figure 3 we show the Nelson-Aalen and the naive local constant estimates of the cumulative hazard for both male and female. Note that for any given t, the estimation of the cumulative hazard provided by the naive local constant estimator is taking into account all occurrences that took place in some period $[t - b, t + b]$ around t. For example, note that for both male and female the most important increase in the number of deaths occurs approximately around the end of the second year after operation or the beginning of the third year, approximately when $t = 621$ days for male and $t = 817$ days for female. This fact is reflected in the corresponding estimates of the naive local constant estimator prior to these time points, providing larger estimates than the Nelson-Aalen estimator.

The ratio between the naive local constant and the Nelson-Aalen estimator, see Figure 4, can be used for comparative purposes. Note that the ratio is quite large for small t's but it decreases for larger t's. This ratio becomes very close to one after day 2000. After day 3000, approximately, the estimates seem to be equal, thus the ratio is 1.

For males, when $t < 779$ days, the naive local constant estimator provides larger estimates than the Nelson-Aalen estimator, except during a short period between day 210 and day 351. The reason is that it includes some information about what is going to happen after the time point being considered, so the substantial increase in the number of deaths occurring between day 621 and day 793 is taken into account. During a second period between day 779 and day 1892, the naive local constant estimator provides



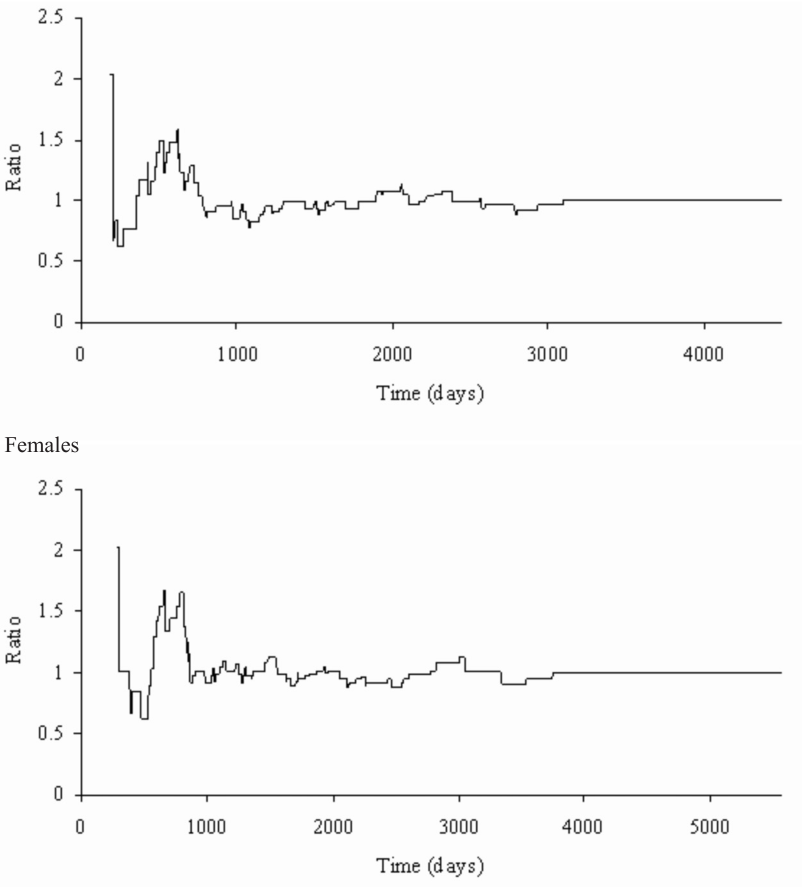
(a) Males

(b) Females

Figure 3: *Comparison between the Nelson-Aalen and the naive local constant estimator for both male and female.*

smaller estimates than the Nelson-Aalen, because it takes into consideration that no sudden increases in the mortality occur in subsequent periods. After day 1892, both estimates are close together, but the cumulative hazard curves do not become equal until day 3091.

A similar pattern is observed for women when comparing both estimates. During the period for $t < 872$ the naive local constant estimator provides larger estimates than the Nelson-Aalen estimator, except between day 386 and day 555. Again the naive local constant estimator captures the important increase that is going to occur in the number of deaths between day 817 and day 872. From $t = 872$ to $t = 2062$, the difference between the naive local constant and the Nelson-Aalen estimates is not so large, but after day 2062 the naive local constant estimator provides smaller estimates than Nelson-Aalen, except during a short period around day 3000. Estimates become equal at day 3745.



(a) Males

(b) Females

Figure 4: *Comparison between the Nelson-Aalen and the naive local constant estimator for both male and female.*

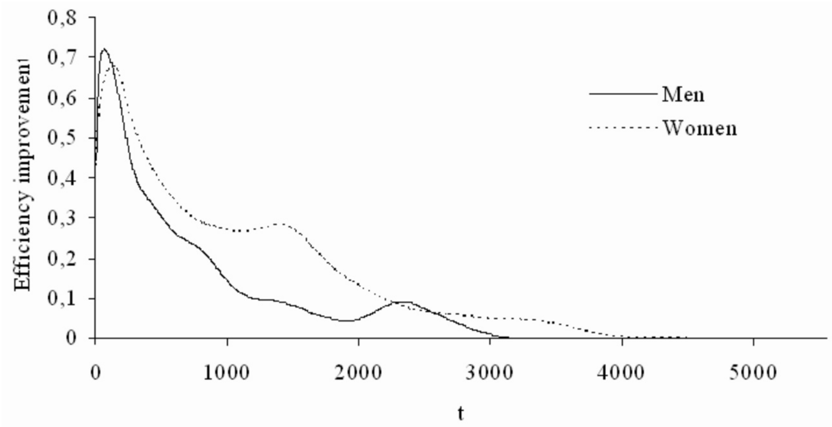


Figure 5: *Relative efficiency gain curve of the naive local constant with respect to the Nelson-Aalen estimator.*

6 Relative efficiency gain

Finally, in Figure 5 we plot the relative efficiency gain curve (1) for male and female. For males, the maximum value is around 72% at $t = 67$ days. For females, the maximum value is around 69% at $t = 167$ days. Note that when $t < 1000$ days the relative efficiency gain is higher than 10% for men and well above 25% for female. Therefore, we conclude that the efficiency improvement of the naive local constant estimator with respect to the Nelson-Aalen estimator is very substantial, especially for small survival times.

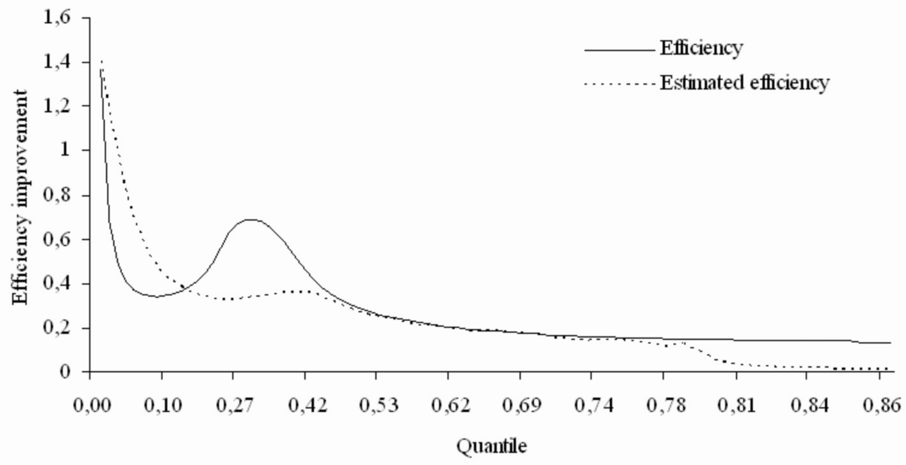


Figure 6: Efficiency improvement for a lognormal distribution $\mu = 1.2$ and $\sigma = 1$.

7 Simulation study

In this section we present a simulation study consisted of generating 1 000 samples of lognormally distributed survival times for 100 individuals. With this survival data, we have estimated the optimal bandwidth parameter and the efficiency with the methodology described in Sections 2 and 4, and we have compared it with the theoretical one. Results are shown in Figure 6. Note that the efficiency improvement is very remarkable specially for short survival times and above 20% in most of the cases. Additionally, it is important to remark that even after estimating the optimal bandwidth parameter, the estimated efficiency gain curve is capturing reasonably well the efficiency gain performance of the naive local constant estimator.

8 Conclusion

We conclude that the naive local constant estimator can be easily applied for estimating the cumulative hazard function based on real survival data. It provides a substantial efficiency improvement with respect to the Nelson-Aalen estimator, especially for small survival times (well above 60%).

Regarding the practical aspects that should be addressed in order to use this estimator, we conclude that it is possible to apply the local linear estimator in order to get an approximation of the hazard and its squared first derivative, which are necessary to estimate the optimal bandwidth. As mentioned before, in case the bandwidth used in the study differs from the optimal one, the efficiency improvement is not the optimal, but it is still positive.

Therefore, we conclude that the naive local constant estimator should be considered as an alternative to the Nelson-Aalen estimator in survival studies. The new estimator can be easily applied and provides with an improved efficiency performance with respect to the Nelson-Aalen estimator.

Acknowledgements

Thanks are given to the Spanish Ministry of Education and Science grant SEJ2004-05052 and SEJ2007-63298.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics*, 6, 701-726.

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

Guillén, M., Nielsen, J. P. & Pérez-Marín, A. (2007). Improving the efficiency of the Nelson-Aalen estimator: the naive local constant estimator. *Scandinavian Journal of Statistics*, 34, 419-431.

Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association*, 53, 457-481.

Klein, J. P. & Moeschberger, M. L. (1997) *Survival Analysis Techniques for Censored and Truncated Data*, Springer Verlag, New York.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics*, 14, 945-965.

Nielsen, J. P. & Tanggaard, C. (2001). Boundary and bias correction in kernel hazard estimation, *Scandinavian Journal of Statistics*, 28, 675-698.

Peña, E. A., Rohatgi, V. K. (1993). Small sample and efficiency results for the Nelson-Aalen estimator. *Journal of Statistical Planning and Inference*, 37, 193-202.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.