Weißbach, Rafael; Gefeller, Olaf

Working Paper

A rule-of-thumb for the variable bandwidth selection in kernel hazard rate estimation

Technical Report // Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, No. 2004,12

Provided in cooperation with:
Technische Universität Dortmund

Suggested citation: Weißbach, Rafael; Gefeller, Olaf (2004) : A rule-of-thumb for the variable bandwidth selection in kernel hazard rate estimation, Technical Report // Universität Dortmund, SFB 475 Komplexitätsreduktion in Multivariaten Datenstrukturen, No. 2004,12, http://hdl.handle.net/10419/49319

A Rule-of-Thumb for the Variable Bandwidth Selection in Kernel Hazard Rate Estimation

Rafael Weißbach† and Olaf Gefeller‡

† Institut für Wirtschafts- und Sozialstatistik
Fachbereich Statistik
Universität Dortmund, 44221 Dortmund, Germany
E-Mail: Rafael.Weissbach@uni-dortmund.de
Tel: +49-231-755-5419
Fax: +49-231-755-5284

‡Institut für Medizininformatik, Biometrie und Epidemiologie
Medizinische Fakultät
Friedrich-Alexander Universität Erlangen-Nürnberg
Waldstr. 6, 91054 Erlangen, Germany
E-Mail: Olaf.Gefeller@rzmail.uni-erlangen.de
Tel: +49-9131-85-22750
Fax: +49-9131-85-25740

February 9, 2004

Abstract

In nonparametric curve estimation the decision about the type of smoothing parameter is critical for the practical performance. The nearest neighbor bandwidth as introduced by Gefeller and Dette 1992 for censored data in survival analysis is specified by one parameter, namely the number of nearest neighbors. Bandwidth selection in this setting is rarely investigated although not linked closely to the frequently studied fixed bandwidth. We introduce a selection algorithm in the hazard rate estimation context. The approach uses a newly developed link to the fixed bandwidth which identifies the variable bandwidth as additional smoothing step. The procedure gains further data-adaption after fixed bandwidth smoothing. Assessment by a Monte Carlo simulation and a clinical example demonstrate the practical relevance of the findings.

1 Introduction

The hazard rate is suitable for the assessment any kind of survival data. E.g. in econometrics the behavior of default of obligors to a bank is described by the latter and of interest for risk management purposes. In biometry the post-surgery behaviour of cancer patients can be assess with the hazard rate because of its interpretation as instantaneous failure rate. The hazard rate can quantify the risk of tumor relapse or death as a function of "time after treatment". The assessment can be used for post-surgery care procedures. It can be estimated nonparametrically with the idea of kernel smoothing as in kernel density estimation. An estimate of the cumulative hazard rate is convoluted with a kernel function to estimate the hazard rate. We use the unbiased Nelson-Aalen estimate of the cumulative hazard rate. For an analysis of the Nelson-Aalen estimate derived in counting process theory confer Andersen et al. 1993. As in density estimation the bandwidth is crucial for the performance of the estimate. The bandwidth must not be too big to avoid oversmoothing and systematical bias as well as it must not be too small to avoid the random noise prevent the notion of the underlying structure. Since this balancing problem varies along the time axis with varying density of observations a variable bandwidth is needed. We use the nearest-neighbor bandwidth which automatically adapts for such variability despite its one-dimensional parameter, namely the number of nearest-neighbors. Our aim is to establish a bandwidth procedure that does not parallel the considerations undergone in density estimation but strive for a procedure that directly maps the bandwidth selectors developed for the fixed bandwidth in density estimation.

Such argument enables the use of extensive literature on (optimal) bandwidth selection in density estimation. For an overview of the latter see Jones et al. 1996. As an example we rescale the "normal-scale-rule" by Parzen 1962 which minimizes the asymptotic integrated mean squared error for the kernel estimate of a normal density. We compare the derived method with a fixed bandwidth approach to assess the value of the additional smoothing by the nearest-neighbor bandwidth. Cross-validatory bandwidth selectors for the nearest-neighbor bandwidth definition are the most recent optimality considerations under random censoring, since plug-in procedures are not established. For an overview on cross-validation bandwidth selection see Marron 1987. The comparison of our method with a representative of such cross-validation technics results in encouragement with respect to up-to-date procedures.

2 The basic approach of kernel smoothing in density estimation

Nonparametric functional estimation can be used for a variety of purposes. It can visualize data structure in density estimation. It may also be used for model selection or model checks and it can be employed to identify subgroups in a data set with inhomogeneous behaviour. For the sake of illustrating the basic ideas, let us first consider the simplest example of density estimation via nonparametric kernel smoothing. For the PARZEN-estimator

$$f_n(x) := \int_{\mathbb{R}} \frac{1}{b} K\left(\frac{x-t}{b}\right) dF_n(t)$$

the empirical distribution $F_n(\cdot)$ of the i.i.d. observations $X_1, \ldots, X_n$ with density $f(\cdot)$ is smoothed with the kernel $K(\cdot)$ to get a smooth estimate of $f(\cdot)$. I.e. the observed values' empirical mass is distributed into their neighborhoods of length proportional to b. For given kernel the bandwidth b is the only unknown parameter. It determines for how far each observations' empirical mass of $1/n$ is considered to imply positive density. Or equivalently from how far away from x observations contribute empirical mass to $f_n(x)$. The amount of contribution is determined by the kernel function K but was shown to have little impact on the estimate's quality (Wand and Jones 1995). We use the bi-quadratic kernel, $K(x) = \frac{15}{16}(1-x^2)^2 I_{[-1,1]}(x)$, throughout the article.

3 Choices for the bandwidth

Since in the classical Parzen approach the bandwidth b is a constant, varying numbers of observations are used in the estimation procedure at different points x, namely, many for high density areas or few for small density areas. This leads to the well-known bias-variance trade-off for selection of the fixed bandwidth. To overcome this problem Wagner 1975 used a *variable* bandwidth $R_n^{NN}(\cdot)$ using a constant number of neighbors instead of a constant window width for the density estimation. A formalization of the k^{th}-nearest-neighbor bandwidth with respect to the empirical process is

$$R_n^{NN}(t) := \inf\left\{ r > 0 \mid |F_n(t - \frac{r}{2}) - F_n(t + \frac{r}{2})| \geq \frac{k}{n} \right\}. \tag{1}$$

This clarifies the way the bias-variance trade-off is paid tribute. The estimate of the cumulative distribution function determines the window width and adjusts for smaller bandwidth in large density areas and larger bandwidth for small density areas. This bandwidth – falsely – lead to the definition of the nearest-neighbor density estimate

$$f_n^*(x) := \frac{1}{R_n^{NN}(x)} \int_{\mathbb{R}} K\left(\frac{x - t}{R_n^{NN}(x)} \right) dF_n(t) \tag{2}$$

involving a constant number of neighbors for estimation at each point x. Hence, even for points where the density vanishes estimation results in positive values. Due to that fact $\int f$ is unbounded (Breiman et al. 1977), the estimator itself is not a density and not even a finite measure anymore. Taking the bandwidth as $R_n^{NN}(X_i)$, i.e. depending on the observed values, and inside the integral (2) becomes

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{R_n(X_i)} K\left(\frac{X_i - x}{R_n(X_i)} \right)$$

and guarantees that the estimate will be a probability density (Breiman et al. 1977).
Another problem of the fixed bandwidth b occurs when data are incomplete, e.g. due to (right-)censoring of observations as frequently encountered in survival analysis. Whereas there is no obvious way to account for this situation when choosing a fixed bandwidth, the definition of the nearest-neighbor distance (1) can be extended to censored data by replacing the empirical process F_n by the Kaplan-Meier estimator of the survival function (Kaplan and Meier 1958) as has been suggested by Gefeller and Dette 1992:

$$R_n^{NN}(t) := \inf\left\{ r > 0 \mid |S_n(t - \frac{r}{2}) - S_n(t + \frac{r}{2})| \geq \frac{k}{n} \right\}. \tag{3}$$

The asymptotic behaviour and impact of definiton (3) for nearest-neighbor distances data was investigated in Dette and Gefeller 1995.

4 Kernel hazard rate estimation

It is well known, that any probability distribution can equivalently be parametrized by density or hazard rate. Ideas developed in the context of density estimation can often be transferred to the setting of hazard rates. With respect to kernel estimation of the hazard rate from censored data — instead of smoothing the empirical process F_n to get an estimate of the density F — one can analogously smooth the Nelson-Aalen estimate of the cumulative hazard rate $H(x) = \int_0^x h(t)dt$,

$$H_n(x) = \sum_{i:X_{(i)} \leq x} \frac{\delta_{(i)}}{n-i+1} \tag{4}$$

to obtain an estimate of the hazard rate. Note here, that we restrict ourselves to positive random variables as is the case in survival analysis. Employing the common notation for censored data we define that the n independent observations $X_i = \max\{T_i, C_i\}$, $i = 1, \ldots, n$, refer either to the survival times T_i or the censoring times C_i and that $\delta_i = I_{\{X_i = T_i\}}$ indicate the censoring for $i = 1, \ldots, n$. The order of the censoring indicators $\delta_{(i)}$ is with respect to the corresponding observations $X_{(i)}$. Survival and censoring times are assumed to be stochastically independent. Interest lies in the estimation of the hazard rate $h(\cdot)$ of the survival times T_i.

Combining the estimate (4) with the nearest-neighbor bandwidth (3) the variable kernel estimate for the hazard rate becomes

$$\begin{aligned} h_n(x) &= \int_{\mathbb{R}} \frac{1}{R_n(t)} K\left(\frac{x-t}{R_n(t)}\right) dH_n(t) &(5)\\ &= \sum_{i=1}^{n} \frac{\delta_{(i)}}{n-i+1} \frac{1}{R_n^{NN}(X_{(i)})} K\left(\frac{X_{(i)}-x}{R_n^{NN}(X_{(i)})}\right). &(6) \end{aligned}$$

5 The double-smoothing approach

Now we jointly want to model the variable bandwidths employing the nearest-neighbor idea for uncensored and censored data and even increase generality. To this end, we replace the estimate of the cumulative distribution function in (1) and of the survival function in (3) by a more general monotone stochastic "smoothing process" Ψ_n. We define the generalized bandwidth

$$R_n(t) := \inf\left\{ r > 0 \mid |\Psi_n(t - \frac{r}{2}) - \Psi_n(t + \frac{r}{2})| \geq p_n \right\}. \tag{7}$$

The bandwidth parameter p_n equals the number of nearest neighbors divided by the number of observations in the case of the nearest-neighbor bandwidth. It is crucial to recognize that the fixed bandwidth is included in the generalization. Let

$$\begin{aligned} \Psi_n(\cdot) &= c \cdot id(\cdot) + d \quad \text{and} \\ p_n &= |c| \cdot b \end{aligned} \tag{8}$$

then

$$\begin{aligned} R_n(t) &= \sup\{r > 0 || c \cdot (t - \frac{r}{2}) + d - (c \cdot (t + \frac{r}{2}) + d)| \leq |c| \cdot b\} \\ &= \sup\{r > 0 || - 2c \cdot \frac{r}{2}| \leq |c| \cdot b\} \\ &= \sup\{r > 0 || c| \cdot r = |c| \cdot b\} \\ &\equiv b. \end{aligned}$$

Hence, we view the fixed bandwidth as a generalized bandwidth with respect to a linear *deterministic* smoothing process, a function. This perspective stresses the notion of the fixed bandwidth as a simplification of a more data-adaptive smoothing procedure. Going the opposite direction one can ask for additional data adaptation in smoothing via a two-stage procedure, the "double-smoothing" approach:

Step 1 Determine an appropriate fixed bandwidth b according to an established criterion.

Step 2 Interchange in the generalized bandwidth (7) the linear function by a non-linear stochastic process to allow for further data adaptation.

Step 1: We will now deal with the selection of the fixed b which is not an obvious task. Keep in mind that we want to estimate the hazard rate. Optimal fixed bandwidth selection e.g. with respect to the mean integrated squared error (MISE) is hindered by the fact that this risk does not need to be finite. Consider the case of an exponential distribution, i.e. with constant hazard rate. The MISE for the hazard rate

$$MISE = \int_{\mathbb{R}^+} \frac{h(x)}{1 - G(x)} dx \int_{\mathbb{R}} K^2(z) dz (nb)^{-1} + \frac{1}{4} b^4 \left(\int_{\mathbb{R}} z^2 K(z) dz \right)^2 \int_{\mathbb{R}^+} (h''(x))^2 dx$$

with $G(\cdot)$ as cumulative distribution function for the censoring depends in the absence of censoring asymptotically on $\int_{\mathbb{R}} h(x)dx$ (see Tanner and Wong 1983) which is not bounded for the constant $h(\cdot)$. One possibility to overcome this problem is to restrict the integration

over a meaningful interval. However, it will be difficult to base the decision on that interval's boundaries on rationale ground and the outcome of the MISE will heavily depend upon it which hinders objectivity.

A second way to handle the problem is to include an explicite weight function to focus on the mean integrated weighted squared error (MIWSE) instead of the MISE. One weight function was proposed by Hjort 1991 and transforms the MIWSE for the hazard rate into the MISE for the corresponding density. Another reasoning for the link between bandwidth selectors for the density and the intensity is given in Diggle and Marron 1988. The use of this weight function is linked to our problem of bandwidth selection. The bandwidth should facilitate the use of appropriate counts of observations for estimation at different time points and the distribution of the observation is at first hand governed by the density. As an example we will restrict ourselves to the "normal scale rule" or "Rule-of-Thumb" (see e.g. Silverman 1986). Here the fixed bandwidth is chosen to minimize asymptotically the MISE in kernel density estimation for normal data. The idea dates back to Parzen 1962 but is still appealing for problems where the smoothness of the curve to estimate resembles that of the Gaussian density because of its computational simplicity. The optimal bandwidth is explicitly given by

$$b^{RoT} = \left[\frac{8\pi^{\frac{1}{2}} \int_{\mathbb{R}} K^2(z)dz}{3(\int_{\mathbb{R}} zK^2(z)dz)^2 n} \right]^{\frac{1}{5}} \hat{\sigma} \tag{9}$$

(see Wand and Jones 1995) and thus only depends on kernel-specific constants and on $\hat{\sigma}$, the estimate of the second central moment.

Step 2: Since we motivated the nearest-neighbor bandwidth and are to cope with censoring we will focus on the Kaplan-Meier process to interchange with the linear function. The question is which linear slope c belongs to the fixed bandwidth? In fact, by choosing this slope we can use the bandwidth parameter for the fixed bandwidth $p_n = |c| \cdot b$ (see (8)) as bandwidth parameter k/n for the nearest-neighbor bandwidth. From the notion of the linear smoothing function to be a simplified stochastic process in (7), it becomes evident to view the linear function as an approximation of the cumulative distribution function. A straight-forward estimate of this approximation is a linear regression through the points of the Kaplan-Meier survival function estimation $(X_i, \frac{1}{2}(S_n(X_i-)+S_n(X_i)))$. Note, that the use of the cumulative distribution function of the survival function is interchangeable since only $|c|$ is needed. We will denote the empirical regression coefficient $\hat{\beta}$. The number of nearest neighbors for any fixed bandwidth b is then

$$k_n = \left[n \cdot |\hat{\beta}| \cdot b \right], \tag{10}$$

where the Gaussian brackets $[\cdot]$ are taken to warrant k_n to be a natural number.

Because of the availability of the KAPLAN-MEIER method as well as the linear regression in all common statistical software packages it is obsolete to give an explicit representation of k_n directly depending on the X_i's themselves.

Let us summarize the outcome of the double-smoothing approach for hazard rate estimation: The reasoning suggests the use of the *Rule-of-Thumb* number of

$$k_n = \left[n \cdot |\hat{\beta}| \cdot b^{RoT} \right] \tag{11}$$

nearest neighbors in the definition of the bandwidth (3) with $\hat{\beta}$ denoting the estimate of the slope parameter in a simple linear regression model built upon the (mid-)points of the Kaplan-Meier survival estimate and b^{RoT} defined in (9). This yields the following hazard rate estimate:

$$\hat{h}_{R_n^{NN}}(x) \;:=\; \sum_{i=1}^{n} \frac{\delta_{(i)}}{n-i+1} \frac{1}{R_n^{NN}(X_i)} K\left(\frac{X_i - x}{R_n^{NN}(X_i)} \right). \tag{12}$$

The strong consistency of that estimator is implied in the more general proof in Pflüger and Gefeller 2000.

6 Simulation study

For a simulation one has to select a parametric family for the comparison of true and estimated hazard rates. Natural shapes for modeling hazards of biological and biomedical processes are often relatively simple, e.g. the typical hazard rate of survival after surgery of severe tumors is often unimodal. We chose to use the exponentiated Weibull family (Mudholkar et al. 1995) modeling unimodal, bath-tube, increasing and decreasing shapes including the constant hazard rate. It is defined in terms of the survival function

$$S(x) = 1 - (1 - \exp(-(x/\sigma)^{\alpha}))^{\theta},$$

with $0 < x < \infty$, $\alpha > 0$, $\theta > 0$ and $\sigma > 0$. So that the family is a generalization of the Weibull distribution, which is maintained for $\theta = 1$ (e.g. Mudholkar et al. 1995). In terms of the hazard rate this means:

$$h(x) = \frac{\alpha\theta(1 - \exp(-(x/\sigma)^{\alpha}))^{\theta-1} \exp(-(x/\sigma)^{\alpha})(x/\sigma)^{\alpha-1}}{\sigma(1 - (1 - \exp(-(x/\sigma)^{\alpha}))^{\theta})}. \tag{13}$$

Table 1: Exponentiated Weibull hazard types and shapes

type	hazard rate	parameter space
I	bath-tube	$\alpha > 1$ and $\alpha \cdot \theta < 1$
II	unimodal	$\alpha < 1$ and $\alpha \cdot \theta > 1$
III	monotone decreasing	$\alpha \leq 1$ and $\alpha \cdot \theta \leq 1$
IV	monotone increasing	$\alpha \geq 1$ and $\alpha \cdot \theta \geq 1$

Analysing Formula (13) reveals the four shapes of the hazard rate to be attributable to partitions of the parameter space with limiting lines $\alpha = 1$ and $\alpha \cdot \theta = 1$. The mapping between partitions and shapes is given in Table 1.

Representatives of the four types are chosen such that $[F^{-1}(0.1), F^{-1}(0.9)]$, i.e. the inner 80% areas, are comparable. A simulation study for the four types was conducted to assess the performance of nearest neighbor bandwidth (3) with the number of nearest neighbors according to the Rule-of-Thumb (11). Before we get to aggregated measures from the simulation let us consider two examples. The first one-sample simulation trial in Figure 1 is the estimate of a hazard rate derived from 300 observations under 40%, i.e. heavy, censoring. This practically typical example with moderate sample size demonstrates the correct estimate of the hazard rate in overall trend and absolute magnitude but teaches to be uncertain about estimation of local and overall minima and maxima. This example was chosen to be free of boundary effect to avoid confusion of effects.

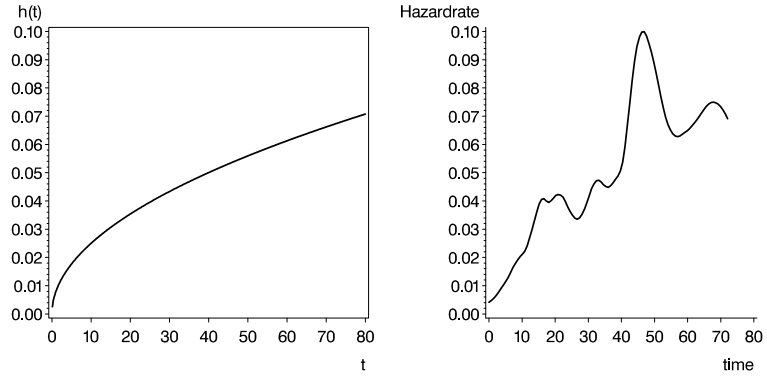

Figure 1. *Monotone increasing hazard rate with $\alpha = 1.5$, $\theta = 1.0$, $\sigma = 33$ true (left side) versus estimate (with $k_n^{RoT} = 64$) (right side)*

The second example in Figure 2 for 300 observations under 40% censoring for the bath-tube shaped exponentiated Weibull hazard reveals the boundary effect which result in two

additional modes at the left and right hand side. Note the upper 10%-quantile $F^{-1}(0.9))$ here is 84.4 and the lower 10%-quantile $F^{-1}(0.1))$ is 1. This advocates to restrict the interpretation on the mentioned inner 80% area.

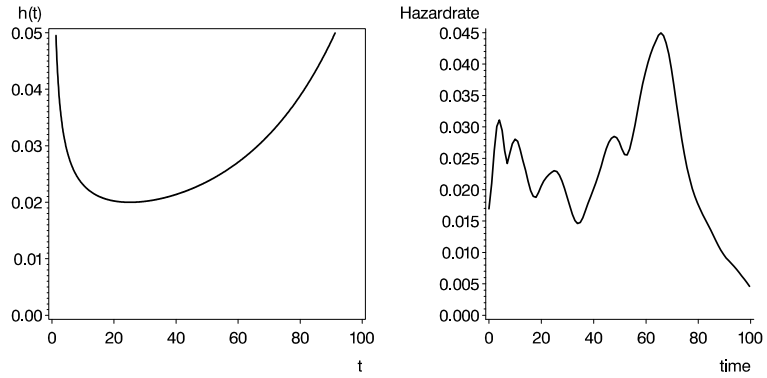

Figure 2. *Bath-tube hazard with* $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$
true (left side) versus estimate (with $k_n^{RoT} = 59$*) (right side)*

We keep that in mind now interpreting the averaged hazard rate estimates over 250 simulation trials. The number of observations is again 300 subject to 40% censoring. The type of hazard rate has shown to have little impact on the quality of the estimation such that for the sake of brevity we restrict the presentation to the type IV bath-tube hazard rates which are most difficult to estimate because of the steep increases at both ends of the support. Figure 3 shows the performance of the derived Rule-of-Thumb for selecting the number of nearest neighbors. A positive bias is seen on starting from time 20 and leaping over into a boundary fading at time 80. The good fit between time 1 and 20 is also due to the decreasing density of that distribution which leads to many observations in that area but is still remarkable since steep increases are hard to detect for nonparametric kernel estimators in general.

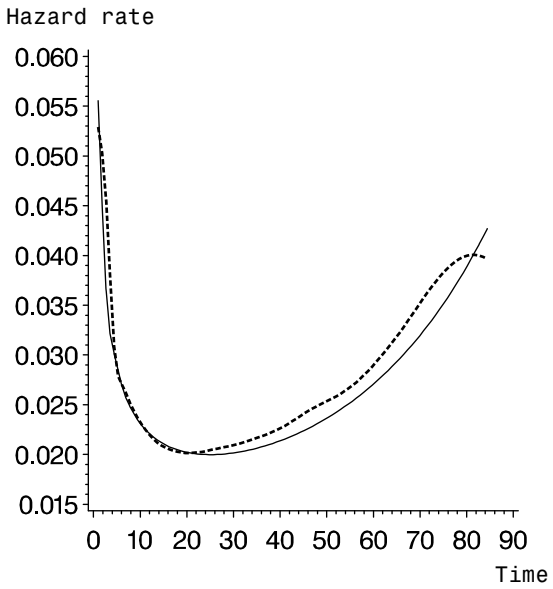

Figure 3. *True exponentiated Weibull hazard rate with* $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ *(solid line) and average of* 250 *simulation trials with estimation from* 300 *observations under* 40% *censoring each with RoT-number of nearest neighbors (dotted line)*

The benefit from the double-smoothing can be seen by comparison with the hazard rate estimate with fixed bandwidth and its Rule-of-Thumb (9) underlying that of the nearest neighbor Rule-of-Thumb (11). The left graph in Figure 4 shows the performance of the latter estimate and reveals a strong bias on the left edge. Such boundary effect for the fixed bandwidth are known and can be taken into account by boundary modifications of the kernel as in Gasser et al. 1985. We do not consider such boundary kernels here for two reasons. At first the boundary is often not known. The origin is a clear boundary in hazard rate estimation but a right side boundary is usually unknown. In density estimation the problem to detect boundaries is even more severe. The second reason is that we want to demonstrate the ability of the nearest neighbor bandwidth to cope with boundary biases even without the knowledge of the exact location and without the additional effort of kernel corrections.

So far bandwidth selection for the nearest neighbor bandwidth in hazard rate estimation was implemented directly e.g. by cross-validation as in Gefeller et al. 1996, where a modified likelihood was maximized to give an optimal bandwidth minimizing asymptotically the expected Kullback-Leibler loss (see Figure 4, right side). The comparison of our Rule-of-Thumb with the latter bandwidth selector detects clear superiority of the Rule-of-Thumb because of the bias towards the origin of the modified likelihood method.

Additionally, the Rule-of-Thumb for the nearest neighbor bandwidth inherits the low variability of the Rule-of-Thumb for the fixed bandwidth. The latter is caused by the fact that data's impact on the bandwidth (9) is restricted to the estimate of the variance. The low variability constitutes a strong advantage over the cross-validation method because cross-validation methods are known to result in highly variable bandwidth selectors (Hall et al. 1987). Another advantage is the computational effort in comparison with cross-validation methods. Especially the modified likelihood maximization with respect to the number of nearest neighbors exacts the complete enumeration over all possible n numbers.

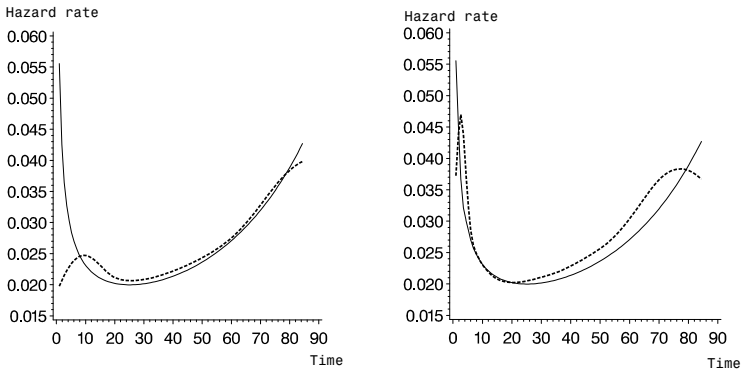

Figure 4. *True exponentiated Weibull hazard rate with* $\alpha = 5$, $\theta = 0.1$, $\sigma = 100$ *(solid line) and average of* 250 *simulation trials with estimation from* 300 *observations under* 40% *censoring each (dotted line) with fixed bandwidth and Rule-of-Thumb (left) and nearest neighbor bandwidth and modified likelihood cross-validation (right)*

7 A clinical example

In the practical example of a bladder cancer study (Siu et al. 1998) one task was to determine whether a low percentage of staining for Metallothionein in the tissue of the cancer predicts for longer survival or lower instantaneous risks of death after surgery and under chemotherapy. 114 patients were stratified into two groups of 45 patients with enriched Metallothionein (of more than 10% of the tumor tissue cells with positive staining) and 69 patients with less than or equal to 10%. Two patients of the first and 15 of the second group were censored.

A standard procedure in the descriptive analysis of censored survival times are Kaplan-Meier survival curves estimates. The two estimates are depicted in Figure 5. [1] The

[1]The survival curve estimation was realized using the SAS procedure "lifetest".

crossing of the two curves at about 500 days indicates a change in instantaneous risk at an earlier point of time. But it leaves the question open where the risk tendencies change. Besides the quantitative information no further insight can be gained from the curves. Noting that after 1000 days only seven uncensored observations occur strongly suggests to restrict the interpretation also for the hazard rate estimate up to that point.

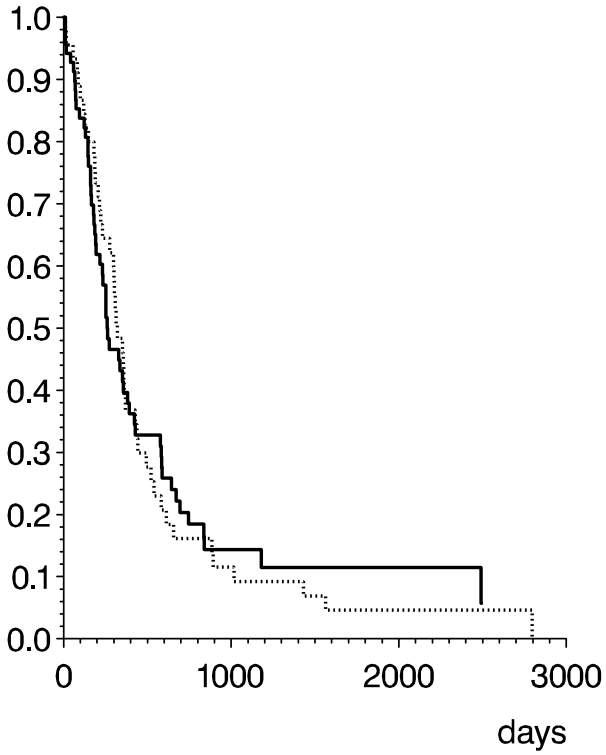

Figure 5. *Survival probabilities according to* Kaplan-Meier *for Metallothionein Strata of* $\leq 10\%$ *(full line) and* $> 10\%$ *staining (dotted line)*

The hazard rate estimates with the developed Rule-of-Thumb number of nearest neighbors in the bandwidth are shown in Figure 6 and give an impression on the changing point of risk superiority-inferiority. After 300 days the lack of Metallothionein seems to result in the lower risk which was the model-based hypothesis derived from molecular biology. Metallothionein is suspected to eject the poison from the chemotherapy out of the cells and even the tumor cells such that the therapy may tend to fail. Death due to the surgery may be the dominant cause of failure before that point of time. As indicated by the simulation study, the observation number – although only subject to censoring of around 10% jointly – indicates that interpretation of the modes and inter-sectional points has to be done carefully and at most for generating hypotheses.

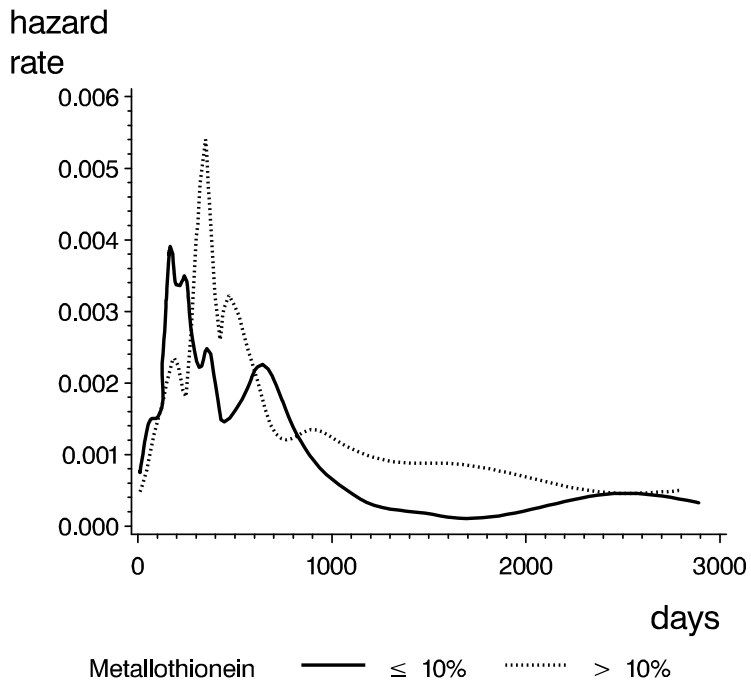

Figure 6. *Hazard rate estimates for Metallothionein Strata with* $k_n^{RoT} = 12$ *for* $\leq 10\%$ *and* $k_n^{RoT} = 17$ *for* $> 10\%$

But besides the care one has to interpret the shapes of the estimates for small sample situations – which is an inherent problem for nonparametric curves estimates – the methodology constitutes a valuable tool to explore censored as well as uncensored data more than only able to compete with the parent histogram still the mostly used nonparametric tool to display the structure of univariate continuous data.

8 Summary

For the kernel hazard rate estimator we advocated for the nearest neighbor bandwidth especially in the survival analytic scenario of censored data. We identified the nearest neighbor bandwidth approach as additional smoothing after fixed bandwidth smoothing which let us to a fast bandwidth selection procedure based on selectors for the fixed bandwidth. We argue that the fixed bandwidth in kernel hazard rate estimation can be selected from density estimation enabling the use of a large amount of literature on optimal bandwidth selection for density estimation. We proved the superiority of our approach over the fixed bandwidth estimation in general and the nearest neighbor bandwidth estimation with cross-validated number selection especially. As an example we demonstrated

the applicability of the bandwidth selector based on the famous "normal scale rule" fixed bandwidth selector for a clinical study.

Acknowledgement: The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

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

L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977.

H. Dette and O. Gefeller. Definitions of nearest neighbour distances for censored data on the nearest neighbour kernel estimators of the hazard rate. *Nonparametric Statistics*, 4:271–282, 1995.

P. Diggle and J.S. Marron. Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association*, 83(403):793–800, 1988.

T. Gasser, H.-G. Müller, and V. Mammitzsch. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, 47:238–352, 1985.

O. Gefeller and H. Dette. Nearest neighbour kernel estimation of the hazard function from censored data. *Journal of Statistical Computation and Simulation*, 43:93–101, 1992.

O. Gefeller, R. Pflüger, and T. Bregenzer. The implementation of a data-driven selection procedure for the smoothing parameter in nonparametric hazard rate estimation using sas/iml software. In *Proceedings of the 13th SAS European Users Group International Conference*. SAS Institute Inc. Carry, SAS Institute Inc. Carry, 1996.

P. Hall, T.C. Hu, and J.S. Marron. On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Annals of Statistics*, 15:163–181, 1987.

N.L. Hjort. Semiparametric estimation of the hazard rates. In *Advanced Study Workshop on Survival Analysis and Related Topics*. NATO, 1991.

M.C. Jones, J.S. Marron, and S.J. Scheather. A brief survey on bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407, 1996.

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

J.S. Marron. A comparison of cross-validation techniques in density estimation. *Annals of Statistics*, 15:152–163, 1987.

G.S. Mudholkar, D.K. Srivastava, and M. Freimer. The exponetial weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, 37:436–445, 1995.

E. Parzen. On the estimation of a probobility density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

R. Pflüger and O. Gefeller. A bridge from the nearest neighbour to the fixed bandwidth in nonparametric functional estimation. In *Proceedings of the* 22^{th} *annual conference of the GfKl*. Gesellschaft für Klassifikation, 2000.

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

L.L. Siu, D. Banerjee, R.J. Khurana, X. Pan, R. Pflüger, I.F. Tannock, and M.J. Moore. The prognostic role of p53, metallothionein, p-glycoprotein, and mib-1 in muscle-invasive urothelial transitional cell carcinoma. *Clinical Cancer Research*, 4:559–565, 1998.

M.A. Tanner and W.H. Wong. The estimation of the hazard function from randomly censored data by the kernel method. *Annals of Statistics*, 11:989–993, 1983.

T.J. Wagner. Nonparametric estimates of probability densities. *IEEE Transactions on Infromation Theory*, 21:438–440, 1975.

M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1995.