

Double-smoothing in Kernel Hazard Rate Estimation

R. Weißbach¹, A. Pfahlberg², O. Gefeller²

¹Institut für Wirtschafts- und Sozialstatistik, Fachbereich Statistik, Universität Dortmund, Dortmund, Germany

²Institut für Medizininformatik, Biometrie und Epidemiologie, Medizinische Fakultät, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Summary

Objectives: In oncological studies, the hazard rate can be used to differentiate subgroups of the study population according to their patterns of survival risk over time. Nonparametric curve estimation has been suggested as an exploratory means of revealing such patterns. The decision about the type of smoothing parameter is critical for performance in practice. In this paper, we study data-adaptive smoothing.

Methods: A decade ago, the nearest-neighbor bandwidth was introduced for censored data in survival analysis. It is specified by one parameter, namely the number of nearest neighbors. Bandwidth selection in this setting has rarely been investigated, although the heuristical advantages over the frequently-studied fixed bandwidth are quite obvious. The asymptotical relationship between the fixed and the nearest-neighbor bandwidth can be used to generate novel approaches.

Results: We develop a new selection algorithm termed *double-smoothing* for the nearest-neighbor bandwidth in hazard rate estimation. Our approach uses a finite sample approximation of the asymptotical relationship between the fixed and nearest-neighbor bandwidth. By so doing, we identify the nearest-neighbor bandwidth as an additional smoothing step and achieve further data-adaptation after fixed bandwidth smoothing. We illustrate the application of the new algorithm in a clinical study and compare the outcome to the traditional fixed bandwidth result, thus demonstrating the practical performance of the technique.

Conclusion: The double-smoothing approach enlarges the methodological repertoire for selecting smoothing parameters in nonparametric hazard rate estimation. The slight increase in computational effort is rewarded with a substantial amount of estimation stability, thus demonstrating the benefit of the technique for biostatistical applications.

Keywords

Disease-free survival, nonparametric statistics, statistical distributions, statistical data interpretation

Methods Inf Med 2008; 47: 167–173

doi:10.3414/ME0447

1. Introduction

The hazard rate is a popular function in biostatistical applications of survival analysis. For example, in assessing post-surgery behavior of cancer patients, it identifies the instantaneous failure rate by quantifying the risk of tumor relapse or death as a function of “time after treatment”. Such an assessment is useful for a rational choice of post-surgery care procedures. The hazard rate can be estimated nonparametrically by applying the concept of kernel smoothing as in kernel density estimation. An estimator of the cumulative hazard rate such as the Nelson-Aalen estimator [1, 2] is convoluted with a kernel function so as to estimate the hazard rate.

As in applications of density estimation, the bandwidth is crucial for the practical performance of the estimator. The bandwidth must not be too large, so as to avoid over-smoothing and systematic bias, but on the other hand, it must not be too small, so as to avoid random noise masking the underlying structure. Since this balancing problem, often referred to as “bias-variance trade-off”, varies along the time axis with a varying density of observations, a variable bandwidth offers clear heuristical advantages. This abstract concept can be put in place with nearest neighbors, which is a general classification concept [3]. The nearest-neighbor bandwidth automatically adapts for such variability, despite being a one-dimensional parameter, namely the number of nearest neighbors.

Our aim is to establish, for our setting, a bandwidth selection algorithm that appropriately maps bandwidth selectors developed for the fixed bandwidth in density estimation. Such a procedure enables the use of

the extensive literature on (optimal) bandwidth selection in density estimation (for an overview of the latter, see [4]). As an example of this strategy, for which we have coined the term “double-smoothing”, we rescale the normal-scale rule by [5], which minimizes the asymptotic integrated mean squared error for the kernel estimator of a normal density.

The paper is structured as follows: in Section 2, we briefly review the principal concepts of kernel smoothing for density (in Section 2.1) and hazard rate (in Section 2.2) estimation with special emphasis on the nearest-neighbor bandwidth. In Section 3, double-smoothing is introduced by generalizing the bandwidth (in Section 3.1) and applying an optimal selection procedure to the fixed bandwidth (in Section 3.2). In Section 4, we demonstrate the application of the method in an oncological study. In the example of a data set covering melanoma patients (with censoring), we identify subgroups uniformly ordered in risk over time.

2. Kernel Smoothing

2.1 Density Estimation

Nonparametric functional estimation can be used for a variety of purposes. In density estimation, it can visualize the intrinsic structure of the data. It can also be used for model selection or model checks and to identify subgroups within a data set. In order to illustrate the basic concepts, let us first consider the simplest example of density estimation via nonparametric kernel smoothing, utilizing the estimator proposed by [5]:

$$\begin{aligned}
 f_n^{fix}(x) &:= \int_{\mathbb{R}} \frac{1}{b} K\left(\frac{x-t}{b}\right) dF_n(t) \\
 &= \frac{1}{nb} \sum_{i=1}^n K\left(\frac{x-X_i}{b}\right) \quad (1)
 \end{aligned}$$

The empirical distribution

$F_n(x) := \frac{1}{n} \sum_{i=1}^n I_{\{X_i, \infty\}}(x)$ of the i.i.d. observations X_1, \dots, X_n with density $f(\cdot)$ is convoluted with the kernel $K(\cdot)$ in order to obtain a smooth estimate of $f(\cdot)$. The second equality is a result of the Stieltjes-integral in the first representation [6]. The empirical mass of each observed value is distributed in its neighborhood with a length b . For a given kernel $K(\cdot)$, the bandwidth b is the only unknown parameter. It determines for which region the empirical mass of each observation of $1/n$ is considered to imply a positive density (or equivalently, at which distance from x , observations contribute empirical mass to $f_n^{fix}(x)$). The amount of contribution is determined by the kernel function $K(\cdot)$, but was shown to exert only a negligible impact on the practical performance of the estimator (see, for example, [6]). We use the popular bi-quadratic kernel, $K(x) = \frac{15}{16} (1 - x^2)^2 I_{[-1,1]}(x)$, throughout the article due to its compact support which is a computational advantage compared to some other kernel functions. In contrast to the Gaussian kernel, the calculation of kernel weights in (1) is only necessary for a small portion of the data points.

In the classical approach the bandwidth b is constant. Thus, varying numbers of obser-

vations are used in the estimation procedure at different points x , namely, many for high-density areas and few for low-density areas. This leads to the well-known bias-variance trade-off in selecting the fixed bandwidth. In order to overcome this problem, [7] introduced 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 the empirical distribution is shown in Figure 1 (Eq. 2).

This clarifies the way the bias-variance trade-off is accounted for. The estimate of the cumulative distribution function determines the window width and adjusts for smaller bandwidth in high-density areas and larger bandwidth for low-density areas. We apply this bandwidth to define the variable density estimator (see Eq. 3 in Fig. 1).

This estimator should not be mixed up with the original nearest-neighbor estimator that used the nearest-neighbor distance around x as bandwidth. The kernel estimator with a fixed bandwidth b and with a nearest-neighbor bandwidth $R_n^{NN}(X_i)$ – but not with a local bandwidth $b(x)$ [8] – result in densities which integrate to one.

A problem with the fixed bandwidth b occurs when data are incomplete, for example, due to the 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 (2) can be extended to censored data by replacing the empirical distribution $F_n(\cdot)$ by the product-

limit estimator of the survival function $S_n(\cdot)$ [9]. The latter is the best nonparametric estimator that is invariant under monotone transformations, as has been suggested by [10] (see Eq. 4 in Fig. 1).

2.2 Hazard Rate Estimation

It is well known that any absolutely continuous probability distribution can be stated equivalently in terms of the density or the hazard rate. Ideas developed in the context of density estimation can thus often be transferred to the setting of hazard rates [11] in survival analysis. We now take account of the typical setting in survival analysis, namely randomly right-censored observations [12]. The i.i.d. sample of survival times T_1, \dots, T_n is no longer observable, because, for each survival time T_i , a random censoring time C_i , $i = 1, \dots, n$, may prevent the observation of T_i if $C_i < T_i$. Survival and censoring times are assumed to be stochastically independent. We observe $X_i = \min\{T_i, C_i\}$, $i = 1, \dots, n$, which are non-negative random variables, i.e. $X_i \geq 0$. Additionally, we know whether an observation is a survival time or a censoring time, that is we observe the censoring indicator $\delta_i = I_{\{X_i = T_i\}}$, $i = 1, \dots, n$. A convention is that the ordering of censoring indicators $\delta_{(i)}$ follows that of the corresponding observations $X_{(i)}$.

The estimation of the hazard rate $h(\cdot)$ of the survival times T_i is of interest. With respect to the kernel estimation of this hazard rate from censored data, instead of smoothing the empirical distribution F_n to obtain an estimate of the density f , one can analogously smooth the Nelson-Aalen estimate of the cumulative hazard rate,

$$H_n(x) := \sum_{i: X_{(i)} \leq x} \frac{\delta_{(i)}}{n - i + 1}, \quad (5)$$

to obtain an estimate of the hazard rate [13]. Combining the estimator with the nearest-neighbor bandwidth (4), the variable kernel estimator for the hazard rate becomes like can be seen in Figure 2 (Eq. 6).

The strong consistency of this estimator is given in [14] and by a more general proof in [15]. We must compare this proposal to the classical estimator with fixed bandwidth, i.e. with

$$\begin{aligned}
 R_n^{NN}(x) &:= \inf \left\{ r > 0 : \left| F_n\left(x - \frac{r}{2}\right) - F_n\left(x + \frac{r}{2}\right) \right| \geq \frac{k}{n} \right\} \quad (2) \\
 f_n^{NN}(x) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{R_n^{NN}(X_i)} K\left(\frac{x - X_i}{R_n^{NN}(X_i)}\right) \quad (3) \\
 R_n^{NN}(x) &:= \inf \left\{ r > 0 : \left| S_n\left(x - \frac{r}{2}\right) - S_n\left(x + \frac{r}{2}\right) \right| \geq \frac{k}{n} \right\} \quad (4)
 \end{aligned}$$

Fig. 1
Equations 2-4

$$h_n^{fix}(x) = \sum_{i=1}^n \frac{\delta_{(i)}}{b(n-i+1)} K\left(\frac{x-X_{(i)}}{b}\right), \quad (7)$$

but first study the bandwidth selection problem for these two estimators.

3. Bandwidth Selection

The aim of this section is to determine an optimal selection of the number of nearest neighbors needed to calculate the hazard rate estimate (6). The selection of the fixed bandwidth needed in (7) is a part thereof.

The idea is to start with a fixed bandwidth. We then increase the bandwidth at loci where data are scarce and decrease the bandwidth where data are dense. This is achieved by multiplying the fixed bandwidth b with the inverse of the density $f(x)$. The bandwidth is now *variable*. Unfortunately, it is unknown, due to the unknown density, but we can estimate $b/f(x)$ by means of the nearest-neighbor bandwidth using its asymptotic property shown under mild assumptions [16],

$$R_n^{NN}(x) \xrightarrow{n \rightarrow \infty} \frac{k}{2nf(x)}. \quad (8)$$

Hence, $R_n^{NN}(x)$ is a consistent estimator for $b/f(x)$, if we set k equal to $2nb$.

There is one disfigurement with this idea. Obviously, because the factor $1/f(x)$ is usually greater than 1, some scaling is needed to ensure that the bandwidth will not generally increase. To this end, we look at the similarity between the fixed and the nearest-neighbor bandwidth from a theoretical perspective.

3.1 Theory

We now define a general bandwidth including (i) the nearest-neighbor bandwidth for censored data and (ii) the fixed bandwidth. In so doing, we replace the distribution function estimator in (4) by a more general monotone stochastic $\Psi_n(\cdot)$, named “smoothing process” (see Eq. 9 in Fig. 3).

In the case of the nearest-neighbor bandwidth, the bandwidth parameter p equals the number of nearest neighbors divided by the

Fig. 2 Equation 6

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

number of observations. The fixed bandwidth is embedded in the generalization with the following choices of $\Psi_n(\cdot)$ and p . For any c and d and

$$\Psi_n(x) := cx + d \text{ and } p := |c|b \quad (10)$$

it is like can be seen in Figure 3 (Eq. 11). Hence, the fixed bandwidth can be interpreted as a generalized bandwidth with respect to a *deterministic* smoothing process, a linear function. The fixed bandwidth, in particular, is a coarse simplification of the nearest-neighbor bandwidth, if we think of the linear function as a linear approximation of the distribution function. In the nearest-neighbor bandwidth, the distribution function is more appropriately estimated from the data, namely by use of the product-limit survival estimator – in the role of $\Psi_n(\cdot)$.

With respect to the scaling problem, instead of using the asymptotic relationship (8) between the k -nearest-neighbors bandwidth $R_n^{NN}(x)$ and $b/f(x)$, we use the link supplied by the generalized bandwidth (9). We will now consider fixed bandwidth smoothing as a plug-in estimation, assum-

ing a uniform pilot distribution. Such a view is justified, because asymptotically, the nearest-neighbor bandwidth does not depend on time x , if and only if, the distribution is uniform.

The question is: what slope c of the linear function is associated with the fixed bandwidth? By choosing the slope, we can use the bandwidth parameter for the fixed bandwidth $p = |c|b$ – see definition (10) – as bandwidth parameter k/n for the nearest-neighbor bandwidth.

The parameter c is the slope of the linear approximation for the distribution function, described by $cx + d$ in (10). In addition to the slope of the linear function, we must specify the intercept d . In survival analysis, the intercept can realistically be assumed as 0, because a positive intercept models the atypical situation of a hazard rate of 0 near the origin.

A linear distribution function, hence a uniform distribution, has a right boundary γ . The relation to the slope is given by $c = 1/\gamma$. We need a specification of γ , including possible censoring. The relation to the mean, being $1/2 \gamma$, facilitates solving the considerably easier first moment estimation. As

$$\begin{aligned} R_n(x) &:= \inf \left\{ r > 0 : \left| \Psi_n\left(x - \frac{r}{2}\right) - \Psi_n\left(x + \frac{r}{2}\right) \right| \geq p \right\} \\ R_n(x) &= \sup \left\{ r > 0 : \left| c\left(x - \frac{r}{2}\right) + d - \left(c\left(x + \frac{r}{2}\right) + d \right) \right| \leq |c|b \right\} \\ &= \sup \left\{ r > 0 : \left| -2c \frac{r}{2} \right| \leq |c|b \right\} = \sup \{ r > 0 \mid |c|r = |c|b \} = b. \end{aligned} \quad (9)$$

Fig. 3 Equations 9 and 11

a generalization of the sample mean for right-censored data, we estimate

$$\hat{\gamma} := 2 \int_{\mathbb{R}_0^+} x dF_n(x)$$

with the Kaplan-Meier estimator $F_n(x) =$

$$1 - S_n(x) = 1 - \prod_{X_{(i)} \leq x} \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)}}.$$

Lebesgues-Stieltjes integration yields

$$\hat{\gamma} = 2 \sum_{j=1}^n X_{(j)} S_n(X_{(j-1)}) \frac{\delta_{(j)}}{n-j+1},$$

with $S_n(X_{(0)}) = 1$.

The number of nearest neighbors corresponding to a fixed bandwidth b is then

$$k := \left\lceil \frac{nb}{\hat{\gamma}} \right\rceil \quad (12)$$

where the Gaussian brackets $\lceil \cdot \rceil$ are taken to ensure that k is an integer.

Due to the two steps involved in the smoothing, i.e. starting with a fixed bandwidth and then adjusting the smoothing by modulating the bandwidth, we coined the term “double-smoothing” for this approach. We apply the approach in the next section to kernel hazard rate estimation under random censoring.

3.2 Example

Many examples of the general procedure are conceivable. We follow [17] in handling the optimal fixed bandwidth selection for the hazard rate estimation. In contrast to minimizing the mean integrated squared error, which must be infinite for hazard rate estimates, we include an explicit weighting function to focus on the mean integrated weighted squared error. The weight function proposed by [17] transforms the mean integrated weighted squared error for the hazard rate into the un-weighted mean integrated squared error for the corresponding density. The use of this weight function is linked to our problem of bandwidth selection. The bandwidth should facilitate the use of an appropriate number of observations for estimation at different time points and the distribution of the observation is governed essentially by the density.

Optimal selection methods for fixed bandwidth density estimation have been identified and implemented in the literature. For a comprehensive overview, see [4]. As an example of an optimal fixed bandwidth selection for density estimation, we apply the normal-scale rule, for which Silverman has introduced the term “rule of thumb” (see e.g. [18]). Specifically, the fixed bandwidth is chosen to minimize asymptotically the mean integrated squared error in kernel density estimation for normal data. The procedure dates back to [5] but is still appealing for problems where the smoothness of the curve to be estimated resembles that of the Gaussian density and because of its computational simplicity. The optimal bandwidth is given explicitly by

$$b^{nsr} = \left[\frac{8\pi^{\frac{1}{2}} \int_{\mathbb{R}} K^2(z) dz}{3 \left(\int_{\mathbb{R}} z^2 K(z) dz \right)^2 n} \right]^{\frac{1}{5}} \hat{\sigma} \quad (13)$$

and only depends on kernel-specific constants, on the sample size n , and on the estimate of the standard deviation $\hat{\sigma}$. For our choice of the bi-quadratic kernel, we have $\int_{\mathbb{R}} K^2(z) dz = \frac{5}{7}$ and $\int_{\mathbb{R}} z^2 K(z) dz = \frac{1}{7}$ [6]. For censored data, we can estimate the variance by

$$\hat{\sigma}^2 := \int_{\mathbb{R}_0^+} (x - \int_{\mathbb{R}_0^+} x d(1 - S_n)(x))^2 d(1 - S_n)(x).$$

The implication for the number of nearest neighbors is obtained by summarizing the outcome of the double-smoothing approach for hazard rate estimation using the fixed bandwidth (13). The normal-scale rule for the fixed bandwidth inserted into (12) suggests the use of

$$k^{nsr} = \left\lceil \frac{nb^{nsr}}{\hat{\gamma}} \right\rceil \quad (14)$$

nearest neighbors for kernel hazard rate estimation using the nearest-neighbor bandwidth. For the fixed bandwidth hazard rate estimator (7), we will use the normal-scale rule (13).

4. Clinical Application

Subsequent to an epidemiological study on risk factors for skin cancer, a follow-up of the case group of this large multi-center

case-control study comprising melanoma patients from seven countries was conducted [19, 20]. Complete information on the vital status was available for 542 patients (= 95% of the original case group) with a median follow-up time of 73 months. Altogether, 184 patients died during the follow-up period, which translates into a degree of censoring of 66%. The primary aim of the investigation entails assessing the effect of different vaccinations on the survival of melanoma patients [21], and the secondary aim consisted of evaluating the prognostic relevance of a variety of tumor-specific characteristics.

For the purposes of illustration, we focus on the role of ulceration of the tumor prior to the surgical removal of the tumor lesion. Previous studies have identified ulceration as one of the most powerful predictors of survival among melanoma patients [22], but temporal changes of mortality risk have not been analyzed so far. In our study sample, 119 patients in the case group had ulcerated tumors prior to surgery. The Kaplan-Meier estimate of this subgroup is consistently lower than that of the subgroup without ulceration (see Fig. 4), and the difference between the two curves is statistically significant with a p-value < 0.0001 for the log-rank test.

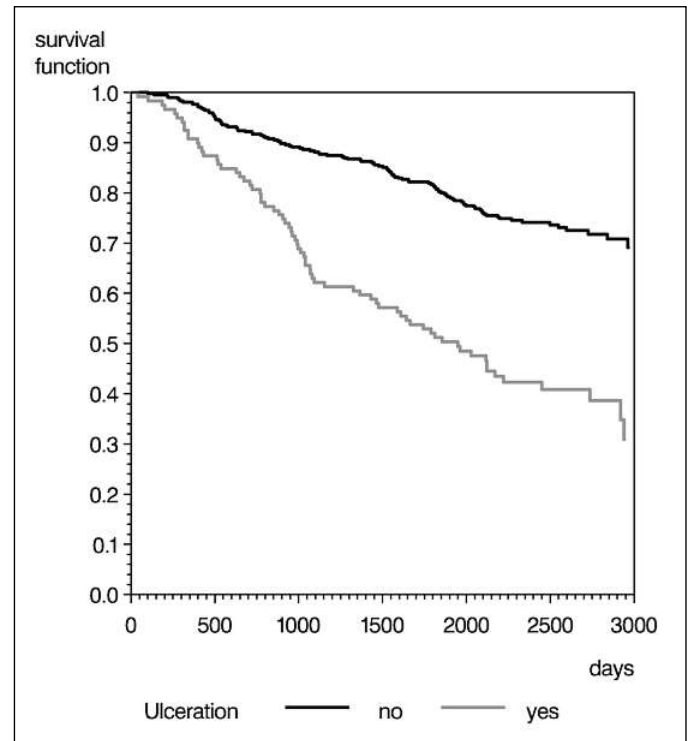
Apart from the finding that the overall risk is different, more detailed information might be desirable. The hazard ratio compares the hazards between the two groups, if we assume the hazard rates to be proportional. According to our data, using the semi-parametric Cox regression, the death hazard for the patients with ulcerated tumors is approximately three times higher than the hazard for the non-ulcerated patients. As a benchmark, the corresponding hazard rates for a Weibull baseline model are indicated in Figure 5. However, the proportional hazards assumption in the Cox regression is critical so that a semi-parametric possibility is to allow for time-dependent regressor coefficients of a particular form [23]. We use the fully nonparametric option of kernel smoothing.

Figure 5 shows the hazard rate estimates of the two groups based on (6). Using rule (14) results in $k^{nsr} = 61$ nearest neighbors in the presence of and in $k^{nsr} = 281$ nearest

neighbors in the absence of ulceration. Additionally, the fixed-bandwidth estimate (7) with normal-scale rule (13) is displayed, normal-scale rule (13) requests a bandwidth of $b^{nsr} = 963$ in the presence and $b^{nsr} = 606$ in the absence of ulceration. Furthermore, the estimate with nearest-neighbor bandwidth chosen by cross-validation is displayed, here the approach results in $k^{cv} = 19$ nearest neighbors in the presence and $k^{cv} = 29$ nearest neighbors in the absence of ulceration. The inclusion of the last two estimates allows the investigation of the properties of the double-smoothing in this example. By looking at the fixed-bandwidth estimate the additional smoothing using (6) can be assessed. Since cross-validation represents an option of directly selecting the number of nearest neighbors in hazard rate estimation, the consequences of replacing the direct by an indirect selection procedure can be studied. The cross-validation approach maximizes the modified likelihood criterion, as discussed by [24] (for computational details of the implementation of this cross-validation approach, see [25]).

From Figure 5, it is evident that, over the entire follow-up period, patients with ulcerated tumors have a higher risk of dying compared to those without ulceration. Even several years after the excision of the tumor, there is still a substantially higher mortality risk for those with originally ulcerated tumors who survived up to that point in time. For the Cox model, this effect is implied by the assumption of a constant hazard ratio; only the nonparametric estimates really allow this interpretation. Another interesting feature that can be seen in the hazard rates refers to the different risk dynamics of the nonparametric estimates over time. There are substantial differences in the assessment. For instance, the cross-validated estimate swings heavily for both groups. This dynamic must be questioned from a medical perspective. As mentioned in [26], cross-validated bandwidth selections yield a high variation, that is, they are highly dependent on the specific data set and under-smoothing – as probably present in the case at hand – occurs frequently. On the other hand, the fixed-bandwidth estimate is not that different from the estimate obtained by double-smoothing. Only the arched shape of

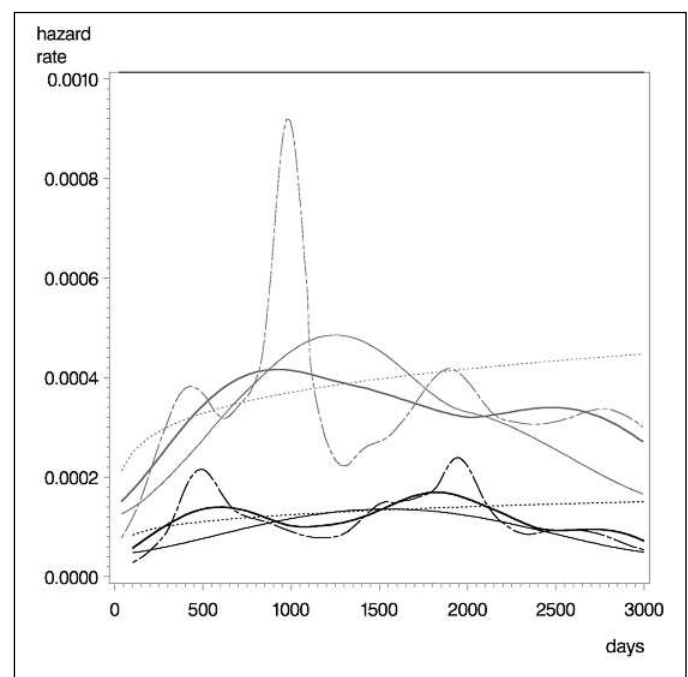
Fig. 4 Kaplan-Meier estimates of the survival function for melanoma patients with and without ulceration of the tumor



the hazard rate for the group without ulceration is disturbing and medically not plausible. The estimate resulting from the double-smoothing approach does not reveal such behavior in this group. The curves for the group with ulceration yield only minor dis-

crepancies. Overall, the hazard rate estimates constructed by using estimator (6) led to smooth curves that did not contradict medical knowledge of typical mortality risk time patterns in melanoma patients during the first years after tumor excision.

Fig. 5 Hazard rate estimates for survival with melanoma in the presence of ulceration (gray lines) and in the absence of ulceration (black lines): i) double-smoothing approach (solid bold lines), ii) fixed-bandwidth estimate using the normal-scale rule (solid thin lines), and iii) nearest-neighbor estimate using cross-validation (dash-dotted lines). The parametric estimates (Weibull) with a constant hazard ratio of three are indicated (dotted lines).



5. Discussion

In this paper, we advocated the nearest-neighbor bandwidth in kernel hazard rate estimation, especially in the survival-analysis scenario of censored data. We demonstrated that the nearest-neighbor bandwidth approach can be regarded as additional smoothing after fixed-bandwidth smoothing. This led us to use the term “double-smoothing” to refer to our procedure which offers a fast-selection algorithm for the number of nearest neighbors based on selectors for the fixed bandwidth. As the actual choice for the fixed bandwidth in our double-smoothing illustration, we used the popular normal-scale rule from [5]. Any alternative choice for a fixed bandwidth, e.g. the use of optimal plug-in bandwidths as discussed in [27] is also possible, but was not considered here, as it does not contribute to the evaluation of the double-smoothing procedure and complicates the selection procedure. It is, however, one of the advantages of our approach that ideas can be drawn from the vast literature on optimal bandwidth selection for density estimation, in order to choose the smoothing parameter in the kernel estimation of the hazard rate.

We observed that, by construction, the nearest-neighbor bandwidth reduces boundary effects as compared to the fixed bandwidth, namely, the bias is reduced. This is especially useful in survival analysis where the origin is a systematic boundary. Although, for the fixed bandwidth, boundary modifications of the kernel function (see e.g. [28]) are a possible means of eradicating the problem, the use of the nearest-neighbor bandwidth offers a solution that reduces the necessity for such modifications.

Cross-validation is a popular method for selecting the smoothing parameter in kernel estimation. The comparison of our double-smoothing method with a cross-validated bandwidth selector of the nearest-neighbor bandwidth was conducted in an extensive simulation study using the exponentiated Weibull family as a benchmark distribution [29]. The results confirmed the known disadvantage of cross-validation methods yielding highly variable bandwidth selec-

tors [26]. Compared to a cross-validated choice, the double-smoothing approach resulted in estimators with similar bias, but decreased variance in most situations. Thus, the findings from the simulation study further support the double-smoothing technique.

In order to assess the practical applicability of our bandwidth selection, we investigated the performance of the method in a clinical study from the oncological field. Notwithstanding the need to interpret the shapes of the estimates for moderate samples very cautiously – an inherent problem for all applications of nonparametric curve estimation – we found that the methodology could easily be implemented in such survival analysis applications. The hazard rates estimated using the double-smoothing methodology displayed the structure of the survival data and yielded useful insights into changes in mortality risk over time. Apart from exploratory statistics further developments are conceivable, especially in computational statistics. The case study suggests that applying the method to survival analysis, for example, may improve the L^1 -Wassermann distance used for survival trees [30].

Acknowledgments

We are indebted to three reviewers for their valuable comments. Furthermore, the financial support of the Deutsche Forschungsgemeinschaft (SFB 475, “Reduction of complexity in multivariate data structures,” and Grant Ge 637/3) is gratefully acknowledged. The EORTC study group of the FEBIM study gave permission for the data from the trial to be used for the purposes of illustration. We thank Professor Kölmel, Göttingen, for this permission. All computations were performed in SAS (Cary, USA). The program code is available from the authors.

References

- Nelson W. Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics* 1972; 14: 945-966.
- Aalen OO. Nonparametric Estimation of Partial Transition Probabilities in Multiple Decrement Models. *Annals of Statistics* 1978; 6: 534-545.
- Bobrowski L. Introduction of Similarity Measures and Medical Diagnosis Support through Separable, Linear Data Transformation. *Methods Inf Med* 2006; 45: 200-203.
- Jones MC, Marron JS, Scheather SJ. A Brief Survey on Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association* 1996; 91: 401-407.
- Parzen E. On the Estimation of a Probability Density Function and the Mode. *Annals of Mathematical Statistics* 1962; 33: 1065-1076.
- Wand MP, Jones MC. *Kernel Smoothing*. London: Chapman & Hall; 1995.
- Wagner TJ. Nonparametric Estimates of Probability Densities. *IEEE Transactions on Information Theory* 1975; 21: 438-440.
- Breiman L, Meisel W, Purcell E. Variable kernel estimates of multivariate densities. *Technometrics* 1977; 19: 135-144.
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; 53: 457-481.
- Gefeller O, Dette H. Nearest Neighbour Kernel Estimation of the Hazard Function from Censored Data. *Journal of Statistical Computation and Simulation* 1992; 43: 93-101.
- Wang JL. In: *Encyclopedia of Biostatistics*. New York: John Wiley & Sons; 1998. pp 4140-4150.
- Lawless JF. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons; 1982.
- Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. New York: Springer; 1993.
- Pflüger R, Gefeller O. A Bridge from the Nearest Neighbour to the Fixed Bandwidth in Nonparametric Functional Estimation, In: R. Decker, W. Gaul (eds). *Classification and Information Processing at the Turn of the Millennium*. Berlin: Springer; 2000. pp 119-126.
- Weißbach R. A general kernel functional estimator with general bandwidth – strong consistency and applications. *Journal of Nonparametric Statistics* 2006; 18: 1-12.
- Dette H, Gefeller O. Definitions of Nearest Neighbour Distances for Censored Data on the Nearest Neighbour Kernel Estimators of the Hazard Rate. *Journal of Nonparametric Statistics* 1995; 4: 271-282.
- Hjort NL. Semiparametric Estimation of the Hazard Rates. In: *Advanced Study Workshop on Survival Analysis and Related Topics*. NATO; 1991.
- Silverman BW. *Density Estimation*. London: Chapman & Hall; 1986.
- Pfahlberg A, Kölmel KF, Grange JM, Mastrangelo G, Krone B, Botev IN, et al. Inverse association between melanoma and previous vaccinations against tuberculosis and smallpox: results of the FEBIM study. *Journal of Investigative Dermatology* 2002; 119: 570-575.
- Kölmel KF, Pfahlberg A, Mastrangelo G, Niin M, Botev IN, Seebacher C, et al. Infections and melanoma risk: results of a multicentre EORTC case-control study. *Melanoma Research* 1999; 9: 511-519.
- Kölmel KF, Grange JM, Krone B, Mastrangelo G, Rossi CR, Henz BM, et al. Prior immunisation with vaccinia or BCG is associated with an improved prognosis of patients with malignant melanoma. An EORTC cohort study on 542 patients. *European Journal of Cancer* 2004; 41: 118-125.

22. Balch CM, Soong SJ, Gershenwald JE, Thompson JF, Reintgen DS, Cascinelli N, et al. Prognostic factors analysis of 17,600 melanoma patients: Validation of the American Joint Committee on Cancer melanoma staging system. *Journal of Clinical Oncology* 2001; 19: 3622-3634.
23. Valenta Z, Pitha J, Podrapska I, Poledne R. Gaining Insight from Flexible Models. *Methods Inf Med* 2006; 45: 186-190.
24. Tanner MA, Wong WH. The estimation of the hazard function from randomly censored data by the kernel method. *Annals of Statistics* 1983; 11: 989-993.
25. Gefeller O, Pflüger R, Bregenzer T. 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. Cary; 1996. pp 1288-1300.
26. Hall P, Hu TC, Marron JS. On the Amount of Noise Inherent in Bandwidth Selection for a Kernel Density Estimator. *Annals of Statistics* 1987; 15: 163-181.
27. Hall P, Sheather SJ, Jones MC, Marron JS. On Optimal Data-Based Bandwidth Selection in Kernel Density Estimation. *Biometrika* 1991; 78: 263-269.
28. Gasser T, Müller HG, Mammitzsch V. Kernels for Nonparametric Curve Estimation. *Journal of the Royal Statistical Society, Series B* 1985; 47: 238-352.
29. Mudholkar GS, Srivastava DK, Freimer M. The Exponential Weibull Family: A Reanalysis of the Bus-Motor-Failure Data. *Technometrics* 1995; 37: 436-445.
30. Radespiel-Tröger M, Gefeller O, Rabenstein T, Hothorn T. Association between Split Selection Instability and Predictive Error in Survival Trees. *Methods Inf Med* 2006; 45: 548-556.

Correspondence to:

Rafael Weißbach
 Institut für Wirtschafts- und Sozialstatistik
 Fachbereich Statistik
 Universität Dortmund
 44221 Dortmund
 Germany
 E-mail: Rafael.Weissbach@uni-dortmund.de