Double one-sided cross-validation of local linear hazards

María Luz Gámiz Pérez

University of Granada, Spain

Enno Mammen

Heidelberg University, Germany and Higher School of Economics, Moscow, Russia

María Dolores Martínez Miranda

University of Granada, Spain

Jens Perch Nielsen

Cass Business School, City University London, U.K.

Summary.
This paper brings together the theory and practice of local linear kernel hazard estimation. Bandwidth selection is fully analysed, including Do-validation that is shown to have good practical and theoretical properties. Insight is provided into the choice of the weighting function in the local linear minimization and it is pointed out that classical weighting sometimes lacks stability. A new semiparametric hazard estimator transforming the survival data before smoothing is introduced and shown to have good practical properties.

Keywords: Aalen's multiplicative model; cross-validation; Do-validation; filtered data; local linear estimation; semiparametric estimation.

1. Introduction

One important practical problem for kernel smoothing applied to survival data is that of finding the optimal level of smoothing. This paper provides a new approach to smoothing optimization for survival data illustrated for one-dimensional local linear kernel hazard estimation. The local linear kernel approach provides a simple intuitive estimator with an elegant solution to the boundary problem. This paper considers a general filtered survival data framework capable of analysing the detailed properties of cross-validation, plug-in and Do-validation bandwidth selectors for local linear kernel hazard estimation. Do-validation is a relatively new bandwidth selection method and the lack of survival data theory on Do-validation could be excused; however, it is about time that the asymptotic theory of classical cross-validation is developed in the important general framework considered in this paper.

The theoretical contributions on kernel smoothing of hazards and smoothing optimization of Patil (1993) and Patil et al. (1994), Jiang and Doksum (2003), Spierdijk (2008), Bagkavos and Patil (2008) and Bagkavos (2011) are all based on more restrictive models, where the survival statistics are approximated by sums of independent identical distributed stochastic variables. Our paper is based on the full filtered counting processes data model rather than on such approximations. The wide literature on right-censored hazard estimation based on individual observations cannot be applied to many central applications in

survival analysis. For example, almost the entire empirical literature on actuarial and demographic mortality prediction and forecasting are based on filtered survival data including left truncation and right censoring. This important class of models is not covered by this literature. Early actuarial work on this kind of discrete survival data dates a long way back as illustrated in Gram (1879,1883) developing local polynomial hazard estimators not far in spirit from our work. More modern expositions considering this kind of discrete survival data in the mathematical statistical literature include Müller et al. (1997), Wang et al. (1998) and Wang (2005).

In this paper we study different implementations of cross-validation and compare them with theoretical plug-in. To our knowledge a practical version of a plug-in hazard estimator has not yet been developed in our general set-up. The finite sample results of this paper favor the Do-validation bandwidth selector to the classical cross-validated one. The performance of the Do-validated bandwidth selector is also impressive compared with theoretical plug-in bandwidth selectors. These findings are in line with recently published finite sample studies recommending Do-validation as a better alternative to feasible plug-in and cross-validation. The original paper on Do-validating density estimators of independent identically distributed (i.i.d.) stochastic variables, Mammen et al. (2011), concludes that infeasible plug-in outperforms Do-validation a little bit in theory and finite sample studies and that plug-in looses its good performance when transferred from infeasible to feasible implementations, at least for kernel density estimation. Further insight for kernel density estimation was added into this discussion by Mammen et al. (2014). Do-validation is perhaps the simplest possible exploitation of indirect cross-validation originally developed by Hart and Lee (2005), Hart and Yi (1998) and Savchuk et al. (2008,2010). Indirect cross-validation transfers the smoothing optimization problem at hand to a more complicated one, where the optimal level of smoothing is easier to obtain. Mammen et al. (2014) develops a class of indirect cross-validation procedures where the limit of the theoretical performance is as good as in infeasible plug-in. In other words, a feasible indirect cross-validation procedure does exist with the same theoretical performance as the infeasible plug-in estimator. One could think of this theoretical optimal and feasible indirect cross-validation procedure as a feasible plug-in estimator. At a first glance this sounds excellent, the problems with plug-in after all came from its practical implementation. It seems that theoretical considerations can only guide us to some extent: when a sufficient level of theoretical excellence has been achieved, then it is the practical performance of the method that counts. In this paper we argue that Do-validation being the simplest and most practical indirect cross-validation method seems to be the best method to use overall. This paper focuses on transferring the simple Do-validation method to survival analysis. Finite sample studies have been considered in the survival density case: Gámiz et al. (2013a) estimated densities on transformed scales. The idea was to develop graphical tests to check if a given frailty model fits well a data set. In another paper, Gámiz et al. (2013b) introduced practical cross-validation and Do-validation to multivariate unstructured hazard estimation. Do-validation showed to have excellent finite sample performance also in this more complicated multivariate framework. The theoretical analyses we provide in this paper was not part of the computational studies Gámiz et al. (2013a,b).

In this paper we study local linear kernel hazard estimation in the full model of filtered stochastic processes: our local linear kernel hazard estimator has a simple structure incorporating otherwise complicated censoring and truncation patterns and immediately lends itself to approximations tailored to the discrete data actually available from many data providers. Old-age mortality estimation is important and challenging as pointed out in for

example Wang et al. (1998). Stability might be hard to obtain and asymptotic theory is not always relevant for the very old-age mortality estimation. This paper provides two new tricks to overcome some of these difficulties and apply them together with Do-validation on real life mortality data. The first trick is on exposure robustness relevant in these very high ages, where exposure might vary a lot from year to year. It turns out that a simple and non-classical choice of weighting in the local linear hazard estimation is sufficient to adjust for exposure instability. The second trick is to develop a semiparametric transformation approach to one-dimensional hazard estimation. Wand et al. (1991) and Bolance et al. (2003) developed such a semiparametric approach in density estimation. Clements et al. (2003), Buch-Larsen et al. (2005) and Gustafsson et al. (2009) introduced equally efficient procedures, where the preliminary transformation of the data is based on a parametric distribution close the considered data. They showed that the better this parametric distribution fits the data, the better performing is the overall semiparametric density estimator. Buch-Kromann et al. (2011) and Jeon and Kim (2013) took advantage of this insight to provide better insurance models for the extreme losses. Such sparse data problems in insurance provide us with similar estimation challenges as old-age mortality. It is therefore natural to develop a semiparametric transformation procedure for hazards. Such a procedure first finds the best possible parametric starting point, then uses it to transform the data, then does the nonparametric estimation including Do-validation on the transformed data and finally transform the estimation results back on the original scale. In our application section this semiparametric estimation procedure provides an immediate method to obtain stable nonparametric old-age mortality estimators.

The paper is organized as follows. In Section 2 the model and the local linear hazard estimator are defined. In Section 3, cross-validation and Do-validation for bandwidth selection of the local linear hazard estimator are introduced. Section 4 provides the asymptotic properties of the bandwidth selectors (details and proofs are deferred to the Appendix). In Section 5 the local linear estimator and its bandwidth selectors are given for the discrete data setting, where only aggregated observations of occurrences and exposures are given. In Section 6 a case study with mortality data is given. In this section we also discuss the choice of the weighting function to achieve exposure robustness. Section 7 includes a finite sample study showing that Do-validation indeed is the preferred bandwidth selector with an excellent practical performance. In Section 8 a new semiparametric version of the local linear hazard estimator based on a transformation procedure of the survival data is introduced and illustrated. All the calculations have been performed with R (R Development Core Team, 2014). An R-package named DOvalidation has been created by the authors (Gámiz et al., 2014), providing original functions that implement all the methods proposed in the paper as well as the datasets used for the empirical illustrations. R-scripts to reproduce the results shown in the paper are also available as supplementary material.

2. The counting process model and the local linear estimator

We observe n individuals, $i = 1, \ldots, n$. Let N_i count observed failures for the ith individual in the time interval $[0, T]$. N_i can take values 0 or 1. We assume that N_i is a one-dimensional counting process with respect to (w.r.t.) an increasing, right continuous, complete filtration $\mathcal{F}_t$, $t \in [0, T]$, i.e. it obeys *less conditions habituelles*, see Andersen et al. (1993) (pp. 60). We assume Aalen's multiplicative model (Aalen, 1978) where the random intensity is written as $\lambda_i(t) = \alpha(t)Y_i(t)$, with no restriction on the functional form of the hazard function

$\alpha(\cdot)$. Again, Y_i is a predictable process taking values in $\{0,1\}$, indicating (by the value 1) when the ith individual is at risk. We assume that $(N_1, Y_1), \ldots, (N_n, Y_n)$ are i.i.d. for the n individuals. Note that this formulation contains the important particular case of a longitudinal study with left truncation and right censoring. In this case we observe tuples (L_i, Z_i, δ_i) $(i = 1, .., n)$ where L_i is the time an individual enters the study, Z_i is the time he/she leaves the study and δ_i is binary and equal to 1 if death is the reason for leaving the study (censoring indicator). Then, the process Y_i above would be $Y_i(t) = I(L_i \leq t < Z_i)$ and $N_i(t) = I(Z_i \leq t)\delta_i$, where $I(\cdot)$ is the indicator function. Hereafter we will work in the general model. Expressions for particular sampling schemes such as left truncation and right censoring can be derived just by substituting the particular expressions of the processes.

Let us consider the above general counting process formulation. An intuitive (ad hoc) hazard estimate is the life table estimate based on grouped lifetimes, which is defined as the ratio between the number of occurrences over the time exposure for the considered groups. A smooth version of this intuitive estimator can be derived using kernel smoothing. Let us define the function $K_b(\cdot)$ as $K_b(\cdot) \equiv b^{-1}K(\cdot/b)$, with a bandwidth parameter $b > 0$ and K being a symmetric probability density function. Then we define the following expressions $O_{LC}(t) = \sum_{i=1}^{n} \int_0^T K_b(t-s)dN_i(s)$ and $E_{LC}(t) = \sum_{i=1}^{n} \int_0^T K_b(t-s)Y_i(s)ds$, which are smooth estimators of the occurrence and integrated exposure, respectively. Therefore a local constant estimator of the hazard function can be defined as the following smoothed occurrence-exposure ratio: $\widehat{\alpha}_{LC}(t) = O_{LC}(t)/E_{LC}(t)$. This estimator simplifies to the number of failures divided by exposure time in a neighborhood of the considered value when a uniform kernel is used. These estimators are related to the popular Ramlau-Hansen estimator written as $\widehat{\alpha}_{RH}(t) = \int_0^T K_b(t-s)d\,\widehat{\Lambda}(s)$, where $\widehat{\Lambda}(s) = \sum_{i=1}^{n} \int_0^s \frac{J(u)}{Y^{(n)}(u)} d\, N_i(u)$ is the Nelson-Aalen estimator of the cumulative hazard (see for example Andersen et al. (1993)), $Y^{(n)}(u) = \sum_{i=1}^{n} Y_i(u)$ is the risk set (also called here the exposure process), which means the number of individuals under observation at time u, and $J(u) = I(Y^{(n)}(u) > 0)$.

Following the local linear approach for hazard estimation developed in Nielsen and Tanggard (2001), we can also construct local linear smoothers of the occurrence and exposure by defining $O_{LL}(t) = \sum_{i=1}^{n} \int_0^T (a_2^K(t) - a_1^K(t)(t-s))K_b(t-s)W(s)dN_i(s)$ and $E_{LL}(t) = \sum_{i=1}^{n} \int_0^T (a_2^K(t) - a_1^K(t)(t-s))K_b(t-s)Y_i(s)W(s)ds$ with $a_j^K(t) = \int_0^T K_b(t-s)(t-s)^j W(s)Y^{(n)}(s)ds, j = 0, 1, 2$. Here, W is a process that is predictable w.r.t. the filtration $\mathcal{F}_t$. We will see that W does not affect the first order asymptotics of the estimator. Dependence of $O_{LL}(t)$, $E_{LL}(t)$ and $a_j^K(t)$ on W will not be indicated in the notation. Two choices of weight functions are of particular interest. The first one is the natural weighting with $W(s) \equiv 1$. The second is the Ramlau-Hansen weighting defined by $W(s) = \{n/Y^{(n)}(s)\}I(Y^{(n)} > 0)$. We will argue in favor of natural weighting in Section 6.

If we consider the ratio of $O_{LL}(t)$ and $E_{LL}(t)$ we get the local linear hazard estimators presented in Nielsen and Tanggard (2001) that is

$$\widehat{\alpha}_{b,K}(t) = \frac{O_{LL}(t)}{E_{LL}(t)}. \tag{1}$$

It can be easily seen that $\widehat{\alpha}_{b,K}(t) = \sum_{i=1}^{n} \int_0^T \bar{K}_{t,b}(t-s)W(s)dN_i(s)$, with

$$\bar{K}_{t,b}(t-s) = \frac{a_2^K(t) - a_1^K(t)(t-s)}{a_0^K(t)a_2^K(t) - \{a_1^K(t)\}^2} K_b(t-s). \tag{2}$$

Notice that $\int_0^T \bar{K}_{t,b}(t-s)W(s)Y^{(n)}(s)ds = 1$, $\int_0^T \bar{K}_{t,b}(t-s)(t-s)W(s)Y^{(n)}(s)ds = 0$, and $\int_0^T \bar{K}_{t,b}(t-s)(t-s)^2W(s)Y^{(n)}(s)ds > 0$, so that $\bar{K}_{t,b}$ can be interpreted as a second order kernel with respect to the measure μ, where $d\mu(s) = W(s)Y^{(n)}(s)ds$. Below we argue that representation (1) helps to explain the behavior of the hazard function especially in the right tail of the distribution, where the exposure tends to be small. We will see that it is useful to display separate plots of the three curves O_{LL}, E_{LL} and $\widehat{\alpha}_{b,K}$. To illustrate this viewpoint, we use the mortality data of Spreeuw et al. (2013) in Section 6.

3. Bandwidth selection by cross-validation and Do-validation

Ramlau-Hansen (1983) suggested cross-validation for kernel hazard estimation using the counting process formulation described above. Recently, the practical papers by Gámiz et al. (2013a,b) have developed cross-validation and the Do-validation method of Mammen et al. (2011) for survival densities and marker dependent hazard estimation. In this paper, cross-validation and Do-validation will be studied for local linear univariate hazard estimation. Let $\widehat{\alpha}_{b,K}$ be a hazard estimator depending on a bandwidth $b > 0$ and a kernel K. Ideally, one would like to choose the smoothing parameter as the minimizer of

$$\Delta_K(b) = n^{-1}\sum_{i=1}^{n}\int_0^T \left\{\widehat{\alpha}_{b,K}(s) - \alpha(s)\right\}^2 Y_i(s)w(s)ds \tag{3}$$

where w is some weight function. In our simulations and our empirical example we always will put $w(s) \equiv 1$.

The minimization of $\Delta_K(b)$ is equivalent to minimizing $n^{-1}\{\sum_{i=1}^n \int_0^T [\widehat{\alpha}_{b,K}(s)]^2 Y_i(s)w(s)$ $ds - 2\sum_{i=1}^n \int_0^T \widehat{\alpha}_{b,K}(s)\alpha(s)\ Y_i(s)w(s)ds\}$. Only the second of these terms depends on the unknown hazard. The cross-validation approach estimates this second term from the data and chooses b as minimizer of

$$\widehat{Q}_K(b) = n^{-1}\left\{\sum_{i=1}^{n}\int_0^T \left[\widehat{\alpha}_{b,K}(s)\right]^2 Y_i(s)w(s)ds - 2\sum_{i=1}^{n}\int_0^T \widehat{\alpha}^{[i]}_{b,K}(s)w(s)dN_i(s)\right\}, \tag{4}$$

where $\widehat{\alpha}^{[i]}_{b,K}(s)$ is the estimator arising when the data set is changed by setting the stochastic process N_i equal to 0 for all $s \in [0,T]$. The cross-validation bandwidth estimate is denoted by $\widehat{b}^K_{CV}$.

Mammen et al. (2011) introduce the Do-validation method by the combination of left- and right-sided cross-validation and because of this it was called Do-validation from (Do)uble cross(-validation) or (D)ouble (o)ne-sided cross(-validation). One-sided cross-validation was previously proposed by Martínez-Miranda et al. (2009) for kernel density estimation. It is based on indirect cross-validation. Indirect cross-validation makes use of the fact that under mild regularity conditions asymptotically optimal bandwidths for two kernel estimators with different kernels K and L differ by a factor that only depends on the two kernels K and L. In indirect cross-validation one applies cross-validation to a kernel estimator with kernel L and afterwards one multiplies the cross-validation bandwidth by the factor depending on K and L to get a bandwidth for the kernel estimator with kernel K. Such a construction makes sense if cross-validation for a kernel estimator with kernel L works better than cross-validation for a kernel estimator with kernel K. It has been shown by asymptotic theory and by simulations that the kernel L can be chosen such that this is indeed the case. In

Do(uble)-validation one takes the average of two indirect cross-validation bandwidths. The two kernels L_1 and L_2 of indirect cross-validation correspond to local linear smoothing with one-sided kernels. They are defined as in (2) but with K replaced by their left-sided or right-sided versions: $K_L(u) = 2K(u)I(u < 0)$ or $K_R(u) = 2K(u)I(u > 0)$, respectively. We denote the resulting local linear kernels by $\bar{K}_{L,t,b}$ or $\bar{K}_{R,t,b}$, respectively. An intuitive reason that cross-validation for kernel estimators with kernel K_L or K_R works better than for K lies in the fact that the asymmetry of the kernels K_L and K_R leads to larger optimal bandwidths. The one-sided cross-validation criteria are given as $\widehat{Q}_{K_L}(b)$ and $\widehat{Q}_{K_R}(b)$, see (4). Finally, the Do-validation bandwidth estimate, $\widehat{b}_{DO}$, is defined as the weighted average

$$\widehat{b}_{DO} = \frac{1}{2}\left(\frac{R(\bar{K}_L^*)}{R(K)}\frac{\mu_2(K)^2}{\mu_2(\bar{K}_L^*)^2}\right)^{1/5}\left(\widehat{b}_{CV}^{K_L} + \widehat{b}_{CV}^{K_R}\right),$$

where $\widehat{b}_{CV}^{K_L}$ and $\widehat{b}_{CV}^{K_R}$ are the minimizers of $\widehat{Q}_{K_L}(\cdot)$ and $\widehat{Q}_{K_R}(\cdot)$, respectively. Here $\bar{K}_L^*$ denotes the equivalent kernel defined in expression (6) in Section 4, and we have defined the functions $\mu_2(L) = \int u^2 L(u)du$ and $R(L) = \int L^2(u)du$, for $L = K, \bar{K}_L^*$.

Note that Do-validation cross-validates twice with two different kernels. Therefore the complexity of the algorithm to compute the Do-validated bandwidth is two times the complexity of estimating standard cross-validation. Since computational complexity of estimating the local linear kernel hazard estimator is already considerable, the computational time to derive the final hazard estimator with Do-validated bandwidth can be challenging. In this paper we solve this computational challenge by discretizing the time scale by a fine grid. This approach was discussed previously by Nielsen and Tanggard (2001) and Gámiz et al. (2013b) and it will be described in Sections 5 and 7.

4. Asymptotic theory

In this section we develop theory for the asymptotic behavior of the Do-validation bandwidth selector. As just explained, the Do-validation bandwidth is the weighted average of two bandwidth selectors based on indirect cross-validation. Our main result contains an asymptotic normality result for the general class of bandwidths that are constructed as weighted averages of bandwidth selectors based on indirect cross-validation. A corollary to this result gives asymptotic normality of Do-validation. Our theorem also includes the case of classical cross-validation. As far as we know, our asymptotic normality results are new even for the classical cross-validation. There is the unpublished thesis work of Nielsen (1990) developing the asymptotic theory of cross-validation and plug-in for the Ramlau-Hansen estimator and the less general, but published, work of Patil (1993) and Patil et al. (1994) developing the same results as Nielsen did, but for i.i.d. right censored data only. The last of these three papers adding the interesting insight that the hazard should indeed be estimated directly and should not be considered as a ratio of two components to be estimated separately. Therefore, our theory closes an important gap in classical smoothing theory and will hopefully facilitate that many more future developments in this important field will be based on the full filtered survival data model, that is so useful for many practical applications. In Mammen et al. (2011) combinations of indirect cross-validation bandwidths were considered for kernel density estimation and asymptotic normality was shown for bandwidth selectors from this class. In this section we show similar results for the case of hazard rate estimation. Although the theoretical results are similar in spirit, the mathematical tools for deriving the results rely on counting process theory and are qualitatively quite

different in nature. The infeasible ideal plug-in estimator is analysed in full detail. Since feasible plug-in procedures have the exact same large sample performance as the infeasible plug-in, therefore our theory includes theory on feasible plug-in methods (including most bootstrap bandwidth selector methods as well, see for example González-Manteiga et al. (1996)).

For a weight function W we consider the local linear estimators $\widehat{\alpha}_{b,L}$ defined as in expression (1) and in the expression above (2) with kernels $L = K$ and $L = L_j$. The pointwise asymptotic properties of $\widehat{\alpha}_{b,L}$ were derived in Theorem 5.1 of Nielsen and Tanggard (2001). Assuming that the kernel L is a symmetric function, it was shown in Nielsen and Tanggard (2001) that

$$(nb)^{1/2}(\widehat{\alpha}_{b,L}(t) - \alpha(t) - b^2 B_t) \to N(0, V_t^2) \text{ in distribution,} \tag{5}$$

where $B_t = \frac{1}{2}\mu_2(L)\alpha''(t)$, $V_t = R(L)\alpha(t)\{\gamma(t)\}^{-1}$, and $\mu_2(\cdot)$, $R(\cdot)$ being the functions of the kernel defined in Section 3. It can be shown that for an asymmetric kernel L the asymptotic result (5) is exactly the same apart from the kernel constants involved. Specifically, these constants become the values $\mu_2(\bar{L}^*)$ and $R(\bar{L}^*)$, involving the equivalent kernel

$$\bar{L}^*(u) = \frac{\mu_2(L) - \mu_1(L)u}{\mu_2(L) - \{\mu_1(L)\}^2} L(u), \tag{6}$$

where $\mu_1(L) = \int uL(u)du$. In Lemma 3 of the Appendix A we state a uniform asymptotic expansion for the Integrated Squared Error (ISE), $\Delta_L(b)$, see (3). We show that the asymptotic integrated squared error is equivalent to $M_L(b)$ where

$$M_L(b) = b^4 \mu_2^2(\bar{L}^*) \int_0^T \left(\frac{\alpha''(t)}{2}\right)^2 \gamma(t)w(t)dt + (nb)^{-1} R(\bar{L}^*) \int_0^T \alpha(t)w(t)dt.$$

These asymptotic ISE expansions lead to the following asymptotically optimal deterministic bandwidth selector for the local linear hazard estimator with kernel L:

$$b_{MISE}^L = C_{0,L} n^{-1/5}, \text{ where } C_{0,L} = \left\{ \frac{R(\bar{L}^*)}{\mu_2^2(\bar{L}^*)} \frac{\int_0^T \alpha(t)w(t)dt}{\int_0^T \alpha''(t)^2 \gamma(t)w(t)dt} \right\}^{1/5}. \tag{7}$$

For our symmetric kernel K we have $\bar{K}^* = K$, and thus $R(\bar{L}^*)$ and $\mu_2(\bar{L}^*)$ can be replaced by $R(K)$ or $\mu_2(K)$, respectively. The ISE-optimal bandwidth $\widehat{b}_{ISE}^L$ is defined as the minimizer of the ISE criterion $\Delta_L(b)$. To simplify the mathematical asymptotic discussion we assume that $\widehat{b}_{ISE}^L$ is defined as minimizer over the interval $I_n^* = [a_1^* n^{-1/5}, a_2^* n^{-1/5}]$ where the constants $a_2^* > a_1^* > 0$ are chosen such that $a_1^* < C_{0,L} < a_2^*$ for $L = K$ and $L = L_j$ with $j = 1, ..., J$. Lemma 3 shows that $\widehat{b}_{ISE}^L = b_{MISE}^L + o_P\left(n^{-1/5}\right)$. As above, the cross-validation selector $\widehat{b}_{CV}^L$ is defined as the minimizer of the cross-validation criterion: $\widehat{Q}_L(b)$, see (4). Again, to simplify the mathematical asymptotic discussion we assume that $\widehat{b}_{CV}^L$ is defined as the minimizer over the interval I_n^*.

In the following theorem we study the asymptotics of weighted combinations of indirect cross-validation selectors:

$$\widehat{b}^* = \sum_{j=1}^J \omega_j \rho_j \widehat{b}_{CV}^{L_j} \quad \text{with } \rho_j = \rho(\bar{L}_j^*) = \left\{ \frac{R(K)\mu_2^2(\bar{L}_j^*)}{\mu_2^2(K)R(\bar{L}_j^*)} \right\}^{1/5}, \tag{8}$$

for kernels L_j and some weights ω_j with $\sum_{j=1}^{J}\omega_j = 1$. Note also that the definition of $\widehat{b}_{DO}$ is of this form because symmetry of K implies that $R(\bar{K}_L^*) = R(\bar{K}_R^*)$, $\mu_2(\bar{K}_L^*) = \mu_2(\bar{K}_R^*)$ and therefore $\rho(\bar{K}_L^*) = \rho(\bar{K}_R^*)$. The following theorem contains our main theoretical result. It states consistency and asymptotic normality of $\widehat{b}^*$.

THEOREM 1. *Under A1-A3, the bandwidth selector* $\widehat{b}^*$ *in (8) satisfies* $n^{3/10}\left(\widehat{b}^* - b_{MISE}^K\right) \to N\left(0,\sigma_1^2\right)$, *and* $n^{3/10}\left(\widehat{b}^* - \widehat{b}_{ISE}^K\right) \to N\left(0,\sigma_2^2\right)$, *in distribution, where*

$$\begin{aligned}
\sigma_1^2 &= S_1 \int \left(\sum_{j=1}^{J}\omega_j \frac{R(K)}{R(\bar{L}_j^*)}[H_{L_j} - G_{L_j}](\rho_j u)\right)^2 du,\\
\sigma_2^2 &= S_2 + S_1 \int \left(\sum_{j=1}^{J}\omega_j \frac{R(K)}{R(\bar{L}_j^*)}[H_{L_j} - G_{L_j}](\rho_j u) - H_K(u)\right)^2 du,\\
S_1 &= \frac{2}{25}\frac{R(K)^{-7/5}\left(\int \alpha^2(t)w^2(t)\ dt\right)}{\mu_2(K)^{6/5}\left(\int \alpha''(t)^2\gamma(t)w(t)\ dt\right)^{3/5}\left(\int \alpha(t)w(t)\ dt\right)^{7/5}},\\
S_2 &= \frac{4}{25}R(K)^{-2/5}\left(\int \alpha(t)w(t)\ dt\right)^{-2/5}\mu_2(K)^{-6/5}\\
&\qquad \times \left(\int \alpha''(t)^2\gamma(t)w(t)\ dt\right)^{-8/5}\left(\int \alpha''(t)^2\gamma(t)w^2(t)\alpha(t)\ dt\right),
\end{aligned}$$

and where, for $L = L_j$ $(j = 1, ..., J)$, *and* $L = K$, *we define* $G_L(w) = I[w \neq 0][\bar{L}^{**}(w) - \bar{L}^{**}(-w)]$ *and* $H_L(w) = I[w \neq 0]\int \bar{L}^*(u)[\bar{L}^{**}(w+u) - \bar{L}^{**}(-w+u)]\ du$ *with*

$$\bar{L}^{**}(u) = -\frac{\mu_2(L) - \mu_1(L)u}{\mu_2(L) - (\mu_1(L))^2}(L(u) + uL'(u)) + \frac{\mu_1(L)u}{\mu_2(L) - (\mu_1(L))^2}L(u).$$

The following corollary immediately follows from Theorem 1. It states consistency and asymptotic normality of the classical cross-validated bandwidth $\widehat{b}_{CV}^K$, the Do-validation bandwidth $\widehat{b}_{DO}$ and the best possible (infeasible) plug-in bandwidth b_{MISE}^K.

COROLLARY 2. *Under A1-A3, the bandwidth selectors* $\widehat{b}_{DO}$, $\widehat{b}_{CV}^K$ *and* b_{MISE}^K *satisfy* $n^{3/10}(\widehat{b}_{DO} - \widehat{b}_{ISE}^K) \to N(0,\sigma_{DO}^2)$, $n^{3/10}(\widehat{b}_{CV}^K - \widehat{b}_{ISE}^K) \to N(0,\sigma_{CV}^2)$ *and* $n^{3/10}(b_{MISE}^K - \widehat{b}_{ISE}^K) \to N(0,\sigma_{MISE}^2)$, *in distribution, where* $\sigma_{DO}^2 = S_2 + S_1\Psi_{K,DO}$, $\sigma_{CV}^2 = S_2 + S_1\Psi_{K,CV}$, *and* $\sigma_{MISE}^2 = S_2 + S_1\Psi_{K,MISE}$, *with* $\Psi_{K,DO} = \int\left(\frac{R(K)}{R(K_L)}[H_{\bar{K}_L} - G_{K_L}](\rho_j u) - H_K(u)\right)^2 du$, $\Psi_{K,CV} = \int [G_K(u)]^2\ du$ *and* $\Psi_{K,MISE} = \int [H_K(u)]^2\ du$.

Corollary 2 follows directly from Theorem 1. The proof of Theorem 1 is given in the Appendix A. All variances consist of two terms. Only the factor $\Psi_{K,\bullet}$ differs for different bandwidth estimators. This factor only depends on the kernels K and $\bar{K}_L^*$. Mammen et al. (2011,2014) provided the exact calculation of this factor for the Epanechnikov kernel and the quartic kernel (in the density case and under the i.i.d. formulation). These calculations can be used to compare the asymptotic performance of the bandwidth selectors. As for kernel density estimation this can also be done for the hazard case. Table 1 shows the values of the

Table 1. Comparison of asymptotic variances among bandwidth selection methods: factor $\Psi_{K,\bullet}$ defined in Corollary 2.

Method	Epanechnikov	Quartic	Sextic
Do-validation	2.19	1.89	2.36
Cross-validation	7.42	5.87	6.99
Plug-in	0.72	0.83	1.18

factor $\Psi_{K,\bullet}$ for the Epanechnikov kernel, the quartic kernel and the sextic kernel. The last kernel has been used in our empirical studies. Plug-in rules achieve the same asymptotic limit as the MISE-optimal bandwidth under appropriate conditions. Thus, the value of $\Psi_{K,\bullet}$ for plug-in rules in Table 1 is identical to $\Psi_{K,MISE}$. When comparing these constants one has to take into account that the term S_2 has to be added to get the value of the asymptotic variance. This makes the difference between Do-validation and plug-in rather small. Do-validation works under weaker conditions than required for plug-in rules. As in the discussion of Do-validation for density estimation this gives strong evidence for a good performance of Do-validation.

5. A discrete formulation in terms of occurrences and exposures

In this section we write the local linear estimator as a function of occurrences and exposures, see (1). Also we will always choose natural weighting $W(s) \equiv 1$. Typically survival data are not provided as continuous data, in contrast to our continuous model. They are given as aggregated numbers. One reason data providers use this data format is to reduce data size. Another reason might be tradition and habit.

The mortality data that we will use are divided into discrete yearly numbers of occurrences and exposures. This data only allow an approximation of the fully continuous filtered model as it is formulated in this paper. However, one can show that the approximation is sufficiently precise to provide a reasonable fit to the continuous model. We now describe a modification of the local linear estimator for discrete data. We suppose that the following aggregated values of occurrences and exposures are available: $O_r = \sum_{i=1}^{n} \int_{X_{r-1}}^{X_r} dN_i(x)$ and $E_r = \sum_{i=1}^{n} \int_{X_{r-1}}^{X_r} Y_i(x)dx$ for $r = 1, \ldots, m$. Here, $X_1, \ldots, X_m$ are some time points. We allow that they are not equidistant. In particular, this may be the case if the data are transformed to another scale for statistical reasons as we will see in Section 7 below. With $\Delta_r = X_r - X_{r-1}$ we can define $Y_r = E_r/\Delta_r$. This is the average number of individuals which are at risk in the interval $[X_{r-1}, X_r)$ for $r = 1, \ldots, m$ and with $X_0 = 0$. The discrete versions of the estimators $O_{LL}(t)$ and $E_{LL}(t)$ can now be defined as $O_{d,LL}(t) = \sum_{r=1}^{m}(a_{d,2}(t) - a_{d,1}(t)(t - X_r^*))K_b(t - X_r^*)O_r$ and $E_{d,LL}(t) = \sum_{r=1}^{m}(a_{d,2}(t) - a_{d,1}(t)(t - X_r^*))K_b(t - X_r^*)E_r$, where $a_{d,j}(t) = \sum_{r=1}^{m} K_b(t - X_r^*)(t - X_r^*)^j E_r$, $j = 0, 1, 2$, and $X_r^* = (X_{r-1} + X_r)/2$, for $r = 1, \ldots, m$. Finally the discrete version of the local linear hazard estimator is given as the ratio $\widehat{\alpha}_{b,d,LL}(t) = \frac{O_{d,LL}(t)}{E_{d,LL}(t)}$. With $\widetilde{\alpha}_{b,d,LL}^{[r]}$ defined as the estimator after replacing O_r by $O_r - 1$, the cross-validation bandwidth can be defined as the minimizer of $\sum_{r=1}^{m} \left(\widetilde{\alpha}_{b,d,LL}(X_r^*)\right)^2 E_r - 2\sum_{r=1}^{m} \widetilde{\alpha}_{b,d,LL}^{[r]}(X_r^*)O_r$.

Hazard estimation from discretized data has previously been considered by Wang et al. (1998) and Tutz and Pritscher (1996) among others. These previous papers are related to our approach above. Specifically Wang et al. (1998) described estimators which would give the same discrete expression as our local linear hazard estimate, but with the Ramlau-

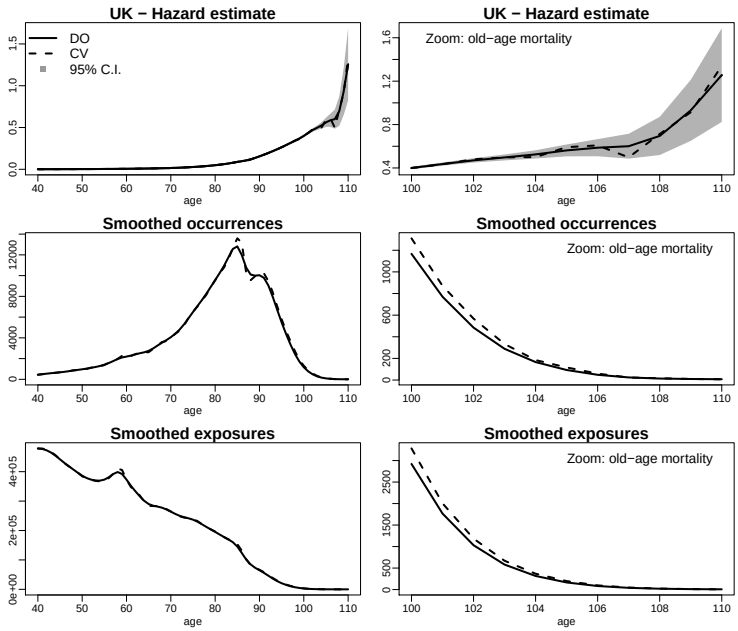

Fig. 1. Case study with mortality data. The top panels show the estimated local linear hazard with 95% confidence bands. The panels below display the two components of the hazard estimators $O_{d,LL}(t)$ and $E_{d,LL}(t)$, defined in Section 5, for United Kingdom. The solid line is obtained using the Do-validation method and the dashed line with cross-validation.

Hansen weighting. Tutz and Pritscher (1996) suggested a discretized version but for the local constant estimator which is the simple Ramlau-Hansen estimator.

6. The local linear hazard estimator in practice

In this section we apply local linear hazard estimation to mortality data of women in the calendar year 2006 from four countries, namely United States, United Kingdom, Denmark and Iceland. The data have also been considered in Spreeuw et al. (2013) and were obtained from the Human Mortality Database. Only the ages from 40 to 110 were included. The data consist of aggregated yearly occurrences and exposures. We have used the discrete version of the local linear estimator described in the last section. Figure 1 displays on the top panel the estimated local linear hazard for United Kingdom. This figure also shows the two components of the estimator, i.e., the smoothed occurrences and the smoothed exposures defined in Section 5respectively. Using the asymptotic properties of the hazard estimator in (5), we have constructed (and plotted in the top panel of Figure 1) 95% confidence intervals for the hazard estimates with Do-validated bandwidth. Table 2 reports the values of the Do-validation and the cross-validation bandwidths for the four countries. The values for Do-validation and cross-validation are quite different. In particular, this is the case for Iceland, where cross-validation leads to an oversmoothed estimation of the hazard curve. This issue will be investigated in Section 7.

Table 2. Case study with mortality data. Estimated bandwidth for each country using the cross-validation and the Do-validation methods.

Country	$\widehat{b}_{CV}$	$\widehat{b}_{DO}$
United States	1.92	3.62
United Kingdom	1.95	4.70
Denmark	7.87	6.43
Iceland	21.41	12.32

Table 3. Old-age mortality data from Iceland. Original data given as occurrences and exposures and the ratio between them.

Age	100	101	102	103	104	105	**106**	107	108	109
Occurrences	6	3	3	1	0	0	**1**	0	0	2
Exposure	11.50	6.83	2.50	1.33	0.50	0.50	**0.17**	0.00	1.00	0.33
Ratio	0.522	0.439	1.20	0.752	0.000	0.000	**5.882**	0	0.000	6.061

We now use the Iceland data to compare natural weights and Ramlau-Hansen weights in local linear smoothing, see the discussion at the end of Section 2. Table 3 gives the observed occurrences/exposures for individuals in Iceland at the ages from 100 to 109. The empirical estimator of the hazard at age 106 is extremely large. The reason is that one death was recorded during this period and that the exposure was only 0.17. This value comes from one 106 years old individual who died in March. The local linear estimators with natural weighting and Ramlau-Hansen weighting are shown in the left plots of Figure 2. We now modify the data and replace the value of exposure for age 106 by 0.005. This very low exposure value would have been reported if the individual would have died in early January instead of in March. The new modified data are shown in Table 4. The resulting new local linear estimators are given on the right hand side of Figure 2. We see that local linear smoothing with natural weights is rather robust. There are nearly no changes of the estimator. On the other hand local linear smoothing with Ramlau-Hansen weights shows drastic changes on the right tails. Note that these changes are caused by a minor change of the data. We argue that this instability also occurs if the number of cases at the boundary is slightly larger. We therefore recommend to use natural weighting instead of Ramlau-Hansen weighting. See also Nielsen and Tanggard (2001) and Nielsen et al. (2009) for more details about this issue.

7. Simulation studies

In this section we compare the finite sample performance of Do-validation bandwidths and cross-validation bandwidth estimates and show that Do-validation corrects some of the

Table 4. A modification in the old-age mortality data from Iceland. The original exposure at the age of 106 has been replaced by the value 0.005 so that the ratio between occurrences and exposures increases dramatically at this age.

Age	100	101	102	103	104	105	**106**	107	108	109
Occurrences	6	3	3	1	0	0	**1**	0	0	2
Exposure	11.50	6.83	2.50	1.33	0.50	0.50	**0.005**	0.00	1.00	0.33
Ratio	0.522	0.439	1.20	0.752	0.000	0.000	**200**	0	0.000	6.061

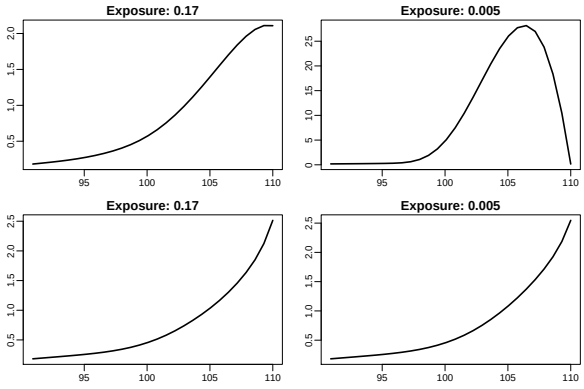

Fig. 2. Exposure robustness analysis of two possible weightings for the local linear hazard estimator: Ramlau-Hansen weighting and natural weighting. The left panels show the hazard estimates from the original old-age mortality data in Iceland (Table 3). The right panels show the hazard estimates obtained from the modified data in Table 4.

well known drawbacks of standard cross-validation such as (i) variability of the bandwidth selector, (ii) bias of the bandwidth selector, and (iii) probability of local minima when selecting the bandwidth, see for example Loader (1999) or Hurvich et al. (1998). We present two different sampling schemes to simulate data from five hazard functions in the next subsections. The first four hazard functions are: $\alpha_1(t) = B(t, 2, 2)$, $\alpha_2(t) = B(t, 4, 4)$, $\alpha_3(t) = 0.6[B(t, 0.5, 0.5) + B(t, 7, 7)]$, $\alpha_4(t) = 0.6[B(t, 0.5, 0.5) + B(t, 4, 2) + B(t, 2, 4)]$. Here $B(t, a, b)$ is the density at t of a Beta distribution with parameters (a, b). These hazard functions have also been used in the simulations in Nielsen and Tanggard (2001). The fifth hazard function is the parametric model estimated by Spreeuw et al. (2013) using the mortality data described in the previous section. This hazard can be written as $\alpha_5(t) = (1 + \sigma^2)^{-1} \exp(a_0 + a_1 t + a_2 t^2) / \int_0^t \exp(a_0 + a_1 s + a_2 s^2) ds$, where $\theta = (a_0, a_1, a_2, \sigma^2)^t$ is a four-dimensional parameter. Here we have considered the maximum likelihood estimates of the parameters calculated from the data corresponding to Iceland (see Spreeuw et al. (2013) for more details). A plot of the five hazard models is provided in the supplementary material. For each hazard datasets are simulated with and without left-truncation.

7.1. Case 1: Without left-truncation

The data are simulated in a discrete grid of time points on the time interval. For models 1 to 4 the time interval is $(0, 1)$ and the grid length $\delta_M = 1/(M + 1)$. For model 5 time is age and it lies in the interval $(40, 110)$ with grid length $\delta_M = 70/(M + 1)$. The grid points are denoted by $\{t_r, r = 1, \ldots, M\}$. Then for a sample of n individuals, failures at time t_r, denoted as O_r are generated from the binomial distribution $Bi\,(Y_r, \alpha_k(t_r)\delta_M)$, for $r = 1, \ldots, M$. Here Y_r denotes the size of the risk set at the beginning of the rth interval of the grid. The total number of simulated occurrences do not sum up to n. Some of the simulated individuals are finally right censored, because they are still at risk at the end of the interval. Therefore our simulated i.i.d. sample contains right censoring. These right censoring rates for the i.i.d. samples are around 20-30% for all models.

Samples are generated with sample sizes $n = 100, 1000$ and 10000 for models 1 to 4 and $n = 50000, 75000$ and 100000 for model 5 (this is comparable to the sample size n=64630 of the mortality dataset). The number of Monte Carlo replications is $R = 1000$. The grid size has been chosen equal to $M = 500$. We have experimented with other choices of M and found that this choice is enough to provide stable results. The local linear hazard estimator has been calculated using the sextic kernel: $K(x) = 3003/2048(1-x^2)^6 I(-1 < x < 1)$. For each model and sample size we compare the performance of the two bandwidth estimates presented in Section 3: the Do-validated bandwidth $\widehat{b}_{DO}$ and the cross-validated bandwidth $\widehat{b}_{CV}$. As benchmarks we calculate two infeasible bandwidths: the ISE-optimal bandwidth $\widehat{b}_{ISE}$ and the MISE-optimal bandwidth $\widehat{b}_{MISE}$. The MISE is approximated by the average of the ISE errors along the R simulated samples. The MISE-optimal bandwidth $\widehat{b}_{MISE}$ is approximated by the bandwidth minimizing this average value. We consider a grid of 100 equispaced bandwidth values around the ISE-optimal bandwidth to compute these four bandwidths. The performance of the bandwidth estimates is analysed with respect to three performance measures, which we denote by m_1, m_2 and m_3, see Table 5. For any $\widehat{b} = \widehat{b}_{DO}, \widehat{b}_{CV}, \widehat{b}_{ISE}, \widehat{b}_{MISE}$, the measure $m_1(\widehat{b})$ denotes the (Monte Carlo estimate of the) MISE of the local linear hazard estimator $\widehat{\alpha}_{\widehat{b},K}$. The bandwidths are compared to the ISE-optimal bandwidth by the measure m_2 which is defined as the average of the differences $\widehat{b} - \widehat{b}_{ISE}$. Thus m_2 is a Monte Carlo estimate of the bias of $\widehat{b}$ with respect to the ISE-optimal bandwidth. Finally we have calculated m_3 which is the standard deviation of the differences $\widehat{b} - \widehat{b}_{ISE}$.

Table 5 shows the simulation results in the above scenarios. Another measure has been added evaluating the relative loss of using Do-validation respectively cross-validation to using the infeasible ISE-optimal bandwidth. The measure is defined as: $Rel\_err = \{m_1(\widehat{b}_{CV}) - m_1(\widehat{b}_{ISE})\}/\{m_1(\widehat{b}_{DO}) - m_1(\widehat{b}_{ISE})\}$. Note that Rel_err indicates when Do-validation outperforms cross-validation considering the criterion m_1. This is the case when this value is above 1. We can see from Table 5, that Do-validation is better than cross-validation for all models and sample sizes. We also see that the improvement from using Do-validation is substantial with a relative error that is often above 2. There is only one out of twelve cases where the result of cross-validation is better than for Do-validation. An inspection of the m_2-values shows that cross-validation has a tendency to undersmooth while Do-validation is slightly oversmoothing. The absolute value of the bias terms were smaller for Do-validation in the first three models and they were smaller for cross-validation in the remaining two. Considering the criterion m_3, we can see that $\widehat{b}_{DO}$ outperforms $\widehat{b}_{CV}$ in almost all cases.

7.2. Case 2: With left-truncation

When adding left truncation L_i only individuals with $L_i \leq Z_i$ are entering the dataset in the simulations and $Y^{(n)}(t)$ is defined as $\sum_{i=1}^{n} I(L_i \leq t \leq Z_i)$, with L_i and Z_i as in Section 2. The left truncation times are simply generated by independent variables from the uniform distribution. Besides these changes, values of exposures and occurrences are generated in the same way as in the previous section. The performance of the bandwidth estimates are again analysed using the same criteria m_1, m_2 and m_3 described above. Also in this more complex model with left truncation Do-validation shows excellent performance properties. For all specifications of the model, Do-validation clearly outperforms cross-validation with Rel_err ranging from 1.27 to 2.75. The complete simulation results in this model are shown in Table 1 in the supplementary material of the paper.

Table 5. Simulation results for datasets without left-truncation. Measure m_1 in columns 3–6 is the empirical MISE for each bandwidth estimate (multiplied by 100 for models 1 to 4 and by 1000 for model 5). The last column shows the relative error Rel_err that compares Do-validation with standard cross-validation. Measures m_2 and m_3 are the average and the standard deviation of the differences $\widehat{b} - \widehat{b}_{ISE}$, respectively.

	Criteria	ISE	$MISE$	CV	DO	Rel_err
Model 1, n =100	m_1	2.499	2.934	4.526	3.314	2.49
	m_2		0.002	-0.005	0.004	
	m_3		0.180	0.315	0.246	
n =1000	m_1	0.370	0.428	0.594	0.447	2.86
	m_2		-0.009	-0.028	-0.016	
	m_3		0.094	0.141	0.104	
n =10000	m_1	0.062	0.067	0.081	0.069	2.71
	m_2		-0.000	-0.013	-0.005	
	m_3		0.043	0.067	0.047	
Model 2, n =100	m_1	3.048	3.629	5.552	4.023	2.57
	m_2		0.002	-0.017	0.006	
	m_3		0.124	0.194	0.150	
n =1000	m_1	0.544	0.609	0.814	0.646	2.66
	m_2		0.001	-0.011	0.003	
	m_3		0.054	0.087	0.066	
n =10000	m_1	0.093	0.100	0.118	0.103	2.50
	m_2		0.000	-0.006	0.000	
	m_3		0.025	0.040	0.029	
Model 3, n =100	m_1	6.370	6.813	9.157	8.287	1.45
	m_2		0.003	0.064	0.133	
	m_3		0.123	0.347	0.407	
n =1000	m_1	1.134	1.192	1.411	1.247	2.45
	m_2		0.003	-0.004	0.008	
	m_3		0.036	0.068	0.048	
n =10000	m_1	0.228	0.233	0.254	0.239	2.36
	m_2		-0.001	-0.001	0.009	
	m_3		0.015	0.027	0.019	
Model 4, n =100	m_1	4.146	4.465	6.766	5.432	2.04
	m_2		0.347	0.090	0.192	
	m_3		0.526	0.943	0.876	
n =1000	m_1	0.803	0.893	1.171	0.967	2.24
	m_2		0.061	0.102	0.079	
	m_3		0.293	0.594	0.477	
n =10000	m_1	0.244	0.254	0.289	0.315	0.63
	m_2		-0.016	0.016	0.147	
	m_3		0.070	0.119	0.105	
Model 5, n =50000	m_1	0.067	0.071	0.082	0.079	1.18
	m_2		-0.095	0.454	-1.024	
	m_3		0.997	1.667	1.192	
n =75000	m_1	0.051	0.054	0.060	0.059	1.11
	m_2		-0.041	0.399	-0.861	
	m_3		0.813	1.370	0.983	
n =100000	m_1	0.041	0.043	0.048	0.047	1.23
	m_2		-0.126	0.341	-0.720	
	m_3		0.740	1.276	0.897	

We have checked cross-validation and Do-validation for the number of local minima in their criterion functions. This has been done by evaluating the criterion on a fine grid of bandwidth values. We have calculated the percentage of times where the score in (4) for the kernel K (cross-validation) or for the one-sided kernels K_L and K_R (Do-validation) have more than one minima on the considered grid of bandwidths. The small number of cases where the Do-validation criterion runs into having several local minima is an indicator for the stability of Do-validation compared to cross-validation. Our results for both models, with and without truncation, show that cross-validation presents multiple local minima in a percentage of cases ranging from 2.0 to 21.2, with a median of 9.8, while Do-validation provides percentages ranging from 0.0 to 2.4, with a median of 0.0. A table with the full summary is provided in Table 2 in the supplementary material of this paper.

8. Hazard estimation for transformed data

In this section we propose a two-step procedure for hazard estimation. First a parametric hazard function $\lambda_i^\theta(t) = \alpha_\theta(t)Y_i(t)$ with $\theta \in \Theta$ is fitted to the data. This parametric fit is used to transform the data in such a way that the underlying hazard would become constant in case the parametric fit indeed would have been the correct underlying model. The parametric fit is in other words used as a kind of prior knowledge to simplify the nonparametric estimation problem. If the prior knowledge is of high quality and the parametric model has good approximation properties, then the resulting nonparametric problem is simplified in the second step. In the second step a local linear hazard estimator is applied to the transformed data. The resulting semiparametric estimator is an alternative to the original fully nonparametric estimator and it is expected to do a better job when the used parametric model is accurate enough.

If the parametric specification α_θ were true then after the time transformation $x = \Lambda_\theta(t)$ with $\Lambda_\theta(t) = \int_0^t \alpha_\theta(s)ds$ the functional form of the hazard on the transformed scale would be simply equal to the standard exponential. That is, the hazard would be equal to one on the transformed scale.

For a given $\theta \in \Theta$ we put $\widetilde{N}_i^\theta = N_i \circ \Lambda_\theta^{-1}$ and $\widetilde{Y}_i^\theta = Y_i \circ \Lambda_\theta^{-1}$. This transformed process follows Aalen's multiplicative hazard model with transformed stochastic hazard $\widetilde{\lambda}_i^\theta(x) = g_\theta(x)Y_i\left(\Lambda_\theta^{-1}(x)\right)$, where $g_\theta(x) = \alpha\left(\Lambda_\theta^{-1}(x)\right)/\alpha_\theta\left(\Lambda_\theta^{-1}(x)\right)$, α is the true hazard, and α_θ is the assumed parametric value. We now carry out our nonparametric local linear smoothing technique on the transformed processes $\widetilde{N}_i^\theta$ and $\widetilde{Y}_i^\theta$ and obtain an estimate $\widehat{g}_\theta$ of g_θ on the transformed time axis. These plots can also be used as a check of the parametric model, see Spreeuw et al. (2013). This is illustrated below by a data example. After back transforming to the original scale we get the semiparametric hazard: $\widehat{\alpha}_\theta(t) = \widehat{g}_\theta\left(\Lambda_\theta(t)\right)\alpha_\theta(t)$. In practice θ is estimated. One could for example use the maximum likelihood estimator $\widehat{\theta}$ of θ, see Borgan (1984).

We now apply the semiparametric transformation method to estimate the hazard function of the mortality data of Spreeuw et al. (2013) discussed above. For simplicity we describe the results for United Kingdom only. As in Spreeuw et al. (2013) we use a parametric mixed hazard model based on a gamma frailty mortality model and we estimate the parameters by a standard maximum likelihood procedure. Then we transform the survival data as just described. The considered parametric mixed mortality model generalizes the classical Gompertz survival model by including more parameters and by including a multiplicative frailty component. The resulting parametric mixed hazard specification of Spreeuw et al.

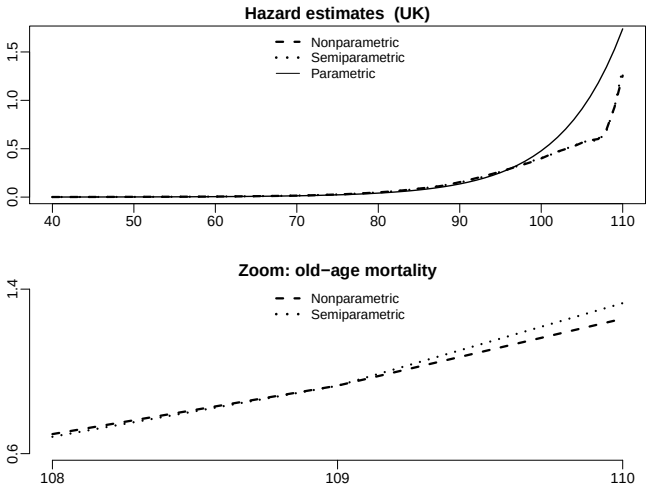

Fig. 3. Hazard estimators with Do-validated bandwidths in the case study for United Kingdom: the parametric estimator, the semiparametric estimator with $\widehat{b}_{DO} = 5.87$; and the nonparametric estimator with $\widehat{b}_{DO} = 4.70$. The bottom panel shows the maximum difference between the semiparametric and the nonparametric estimators at the highest age is about 8% .

(2013) implies that the underlying continuous data come from a counting process with intensity function $\lambda_i^\theta(t) = (1+\sigma^2)^{-1}\exp(a_0 + a_1 t + a_2 t^2)(\int_0^t \exp(a_0 + a_1 s + a_2 s^2)ds)^{-1} Y_i(t) = \alpha_\theta(t) Y_i(t)$, where $\theta = (a_0, a_1, a_2, \sigma^2)^t$ is a four-dimensional parameter.

The final semiparametric approach used in our application was first to calculate the parametric maximum likelihood estimators of θ following Borgan (1984) and then to transform the time axis with the function $\Lambda_{\widehat{\theta}}$, where Λ_θ is the integrated hazard function of the underlying hazard α_θ. Finally, a local linear hazard was estimated to the transformed data and the resulting nonparametric estimator was backtransformed to the original axes. We have compared the resulting hazard estimators from the semiparametric approach with the fully nonparametric local linear estimator considered in Section 2 and the just discussed parametric specification. In general the three estimates are quite close except for the old-age mortality where relative differences up to 8% are found. Semiparametric test theory goes along the lines of many other semiparametric test procedures from survival analysis, see among many others Gandy and Jensen (2005), to evaluate the quality of the parametric transformation approach.

Acknowledgements

The authors gratefully acknowledge the financial support of the Spanish "Ministry of Economy and Competitiveness" by the grants MTM2008-03010 and MTM2013-41383P (European Regional Development Fund), the "Junta de Andalucía" grant TIC-06902 and the European Commission under the Marie Curie Intra-European Fellowship FP7-PEOPLE-2011-IEF Project number 302600. We would also like to thank the Centro de Servicios

de Informática y Redes de Comunicaciones (CSIRC), Universidad de Granada, for providing the computing resources. Support by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953 is gratefully acknowledged. Research of the second author was prepared within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. Finally we thank the Associate Editor and two anonymous reviewers for their constructive comments.

A. Details on the asymptotics and proof of Theorem 1

In this section we describe the assumptions made for the asymptotic theory, introduce additional notation, and provide the proof of the main result.

Assumptions

(A1) *(i) For the expected exposure function* $\gamma(t) = n^{-1}E[Y^{(n)}]$ *it holds that* $\gamma \in C_2([0,T])$*, that it is strictly positive for* $t \in [0,T]$*, and that* $\sup_{s,t\in[0,T],|t-s|\le C_K b} |[Y^{(n)}(t) - Y^{(n)}(s)]/n - [\gamma(t)-\gamma(s)]| = o_P((nb\log n)^{-1/2})$ *and* $\sup_{s\in[0,T]} \left|Y^{(n)}(s)/n - \gamma(s)\right| = o_P((\log n)^{-1})$*, where the constant* C_K *is defined in (A2).*

(ii) The weight W *is a predictable process. There exists a function* $\gamma^* \in C_2([0,T])$*, that is strictly positive for* $t \in [0,T]$*, and that fulfills with the same constant* C_K *as in (i) that* $\sup_{s\in[0,T]} |W(s) - \gamma^*(s)| = o_P((\log n)^{-1})$ *and* $\sup_{s,t\in[0,T],|t-s|\le C_K b}[W(t) - W(s)]/ - [\gamma^*(t) - \gamma^*(s)]| = o_P((nb\log n)^{-1/2})$.

(A2) *The kernels* K *and* L_j $(j = 1,...,J)$ *are compactly supported (i.e., the support is contained in* $[-C_K, C_K]$ *for some constants* $C_K > 0$*). The kernels are continuous on* $\mathbb{R}\backslash\{0\}$ *and have one-sided derivatives that are Hölder continuous on* $\mathbb{R}^- = \{x : x < 0\}$ *and* $\mathbb{R}^+ = \{x : x > 0\}$*, that is there exist constants* c *and* δ *such that* $|g(x)-g(y)| \le c|x-y|^\delta$ *for* $x, y < 0$ *or* $x, y > 0$ *with* g *equal to* K' *or* L_j' $(j = 1,...,J)$*. The left and right-sided derivatives differ at most on a finite set. The kernel* K *is symmetric.*

(A3) *It holds that* $\alpha \in C_2([0,T])$*,* $w \in C_1([0,T])$*. The second derivative of* α *is Hölder continuous with exponent* $\delta > 0$.

Assumptions (A1)–(A3) are rather weak. Assume for instance that $Y_i(t) = \mathrm{I}(L_i \le t < Z_i)$ for some i.i.d. tuples (L_i, Z_i) with joint continuously differentiable density. Then Assumption (A1)(i) can be easily verified with $\gamma(t) = P(L_i \le t < Z_i)$. For the weight $W(s) \equiv 1$ Assumption (A1)(ii) holds trivially. For Ramlau-Hansen weighting $W(s) = \{n/Y^{(n)}(s)\}$ $I(Y^{(n)} > 0)$ it follows from Assumption (1)(i). Assumption (A2) is a weak standard condition on kernels. Assumption (A3) differs from standard smoothness conditions only by the mild additional assumption that the second derivative of the hazard function fulfils a Hölder condition.

We now state and prove some lemmas that we will use in the proof of Theorem 1. For simplicity we assume in the proof that the weight function W is deterministic. If this is not the case W can be treated in the same way as we will do it in the proof for the analysis of $Y^{(n)}(t)/n$. As in the expression above (2) for $L = L_j$ with $j = 1,\dots,J$ and $L = K$ the local linear estimators $\widehat{\alpha}_{b,L}$ are defined as $\widehat{\alpha}_{b,L}(t) = \sum_{i=1}^n \int_0^T \bar{L}_{t,b}(t-s)W(s)\, dN_i(s)$, with $\bar{L}_{t,b}(u)$

defined as in (2). We also define $\alpha^*_{b,L}(t) = \int_0^T \bar{L}_{t,b}(t-s)W(s)\alpha(s)Y^{(n)}(s)ds$. Thus we have that $\widehat{\alpha}_{b,L}(t) - \alpha^*_{b,L}(t) = \sum_{i=1}^n \int_0^T \bar{L}_{t,b}(t-s)W(s)\ dM_i(s) = \int_0^T \bar{L}_{t,b}(t-s)W(s)dM(s)$.

To prove the main result in the paper we first state a uniform asymptotic expansion for the Integrated Squared Error (ISE). Similar expansions are well known for other kernel smoothing estimators but they require some additional work in the hazard case. As outlined before the statement of Theorem 1, the interval $I^*_n = [a^*_1 n^{-1/5}, a^*_2 n^{-1/5}]$ is defined with constants $a^*_2 > a^*_1 > 0$ that fulfill $a^*_1 < C_{0,L} < a^*_2$ for $L = K$ and $L = L_j$ with $j = 1, ..., J$.

LEMMA 3. *Under A1–A3, we get for the ISE, $\Delta_L(b)$, with kernels $L = K$ and $L = L_j$ $(j = 1, \ldots, J)$ that, uniformly for $b \in I^*_n$, $\Delta_L(b) = M_L(b) + o_P\left(n^{-4/5}\right)$.*

PROOF. For brevity we write $\widehat{\alpha} = \widehat{\alpha}_{b,L}$ and $\alpha^* = \alpha^*_{b,L}$. We have that $\Delta_L(b) = n^{-1}\int_0^T \left[\widehat{\alpha}(t) - \alpha^*(t)\right]^2 Y^{(n)}(t)w(t)dt + 2n^{-1}\int_0^T \left[\widehat{\alpha}(t) - \alpha^*(t)\right]\left[\alpha^*(t) - \alpha(t)\right] Y^{(n)}(t)w(t)dt + n^{-1}\int_0^T \left[\alpha^*(t) - \alpha(t)\right]^2 Y^{(n)}(t)w(t)dt$. Below we will apply a martingale central limit theorem. We cannot directly apply this theorem to $\Delta_L(b)$ because the integrand in the definition of $\widehat{\alpha}(t)$ is not predictable. More precisely, the integrand of $\widehat{\alpha}(t)$ contains values of $Y^{(n)}(s)$ with $s > t$ in the terms $a^L_{l,b}(t)$ for $l = 0, 1, 2$. For this reason we use an approximation $\widehat{d^*}(t)$ of $\widehat{\alpha}(t) - \alpha^*(t)$ where $Y^{(n)}(s)$ is replaced by $Y^{(n)}(t) + n\{\gamma(s) - \gamma(t)\}$. We will show that

$$(\log n)^{1/2} n^{1/2} b^{1/2} \left|\widehat{\alpha}(t) - \alpha^*(t) - \widehat{d^*}(t)\right| = o_P(1), \tag{9}$$

uniformly for $0 \le t \le T$ and $b \in I^*_n$, where $\widehat{d^*}(t) = \sum_{i=1}^n \int_0^T \bar{L}^+_{t,b}(t-s)W(s)\ dM_i(s) = \int_0^T \bar{L}^+_{t,b}(t-s)W(s)dM(s)$, $\bar{L}^+_{t,b}(u) = \frac{a^+_{2,b}(t) - a^+_{1,b}(t)u}{a^+_{0,b}(t)a^+_{2,b}(t) - \{a^+_{1,b}(t)\}^2} L_b(u)$, $L_b(u) = b^{-1}L(b^{-1}u)$ and $a^+_{l,b}(t) = \int_0^T L_b(t-s)(t-s)^l W(s)\left[Y^{(n)}(t) + n\{\gamma(s) - \gamma(t)\}\right] ds$ for $j = 1, \ldots, J$ and $l = 0, 1, 2$. Note that the integrand in the definition of $\widehat{d^*}(t)$ is predictable. Expansion (9) follows directly from Assumption (A1). We now apply (9), Assumption (A1) and $\sup_{t\in[0,T]}\left|\widehat{d^*}(t)\right| = O_P((\log n)^{1/2}(nb)^{-1/2})$ and $\sup_{t\in[0,T]}|\alpha^*(t) - \alpha(t)| = O_P((nb)^{-1/2})$. This gives that

$$\begin{aligned}
\Delta_L(b) &= \int_0^T \widehat{d^*}(t)^2\gamma(t)w(t)dt + 2\int_0^T \widehat{d^*}(t)\left[\alpha^*(t) - \alpha(t)\right]\gamma(t)w(t)dt \\
&\qquad + \int_0^T \left[\alpha^*(t) - \alpha(t)\right]^2\gamma(t)w(t)dt + o_P(n^{-4/5}) \\
&= S_{L,1}(b) + S_{L,2}(b) + T_{L,1}(b) + T_{L,2}(b) + o_P(n^{-4/5}),
\end{aligned}$$

uniformly for $b \in I^*_n$, where $S_{L,1}(b) = \int_0^T \bar{H}_{L,b}(u,v)\ dM(u)\ dM(v) - \int_0^T \bar{H}_{L,b}(u,u)\ \alpha(u)\gamma(u)\ du$, $S_{L,2}(b) = 2\int_0^T \delta_{L,b}(u)\ dM(u)$, $T_{L,1}(b) = \int_0^T \bar{H}_{L,b}(u,u)\ \alpha(u)\gamma(u)\ du$, $T_{L,2}(b) = \int_0^T [\alpha^*(u) - \alpha(u)]^2\gamma(u)\ w(u)\ du$, $\bar{H}_{L,b}(u,v) = \int_0^T \bar{L}^+_{t,b}(t-u)\bar{L}^+_{t,b}(t-v)W(u)W(v)\gamma(t)w(t)\ dt$, $\delta_{L,b}(u) = \int_0^T \bar{L}^+_{t,b}(t-u)W(u)\left[\alpha^*(t) - \alpha(t)\right]\gamma(t)w(t)\ dt$.

We now argue that uniformly for $b \in I^*_n$ it holds that $S_{L,1}(b)$ and $S_{L,2}(b)$ are of order $o_P\left(n^{-4/5}\right)$, and that $T_{L,1}(b) = (nb)^{-1}R(\bar{L}^*)\int_0^T \alpha(t)w(t)dt + o_P\left(n^{-4/5}\right)$, $T_{L,2}(b) = b^4\mu_2^2(\bar{L}^*)\int_0^T \left(\frac{\alpha''(t)}{2}\right)^2 \gamma(t)w(t)dt + o_P\left(n^{-4/5}\right)$. We will show the first statement. The second claim follows by similar but simpler arguments. The last two claims can be shown by standard

smoothing theory arguments using that uniformly for $b \in I_n^*$, $C_L b \leq t \leq T - C_L b$ it holds that $R(\bar{L}_t^+) = R(\bar{L}^*) + o(1)$ and $\mu_2(\bar{L}_t^+) = \mu_2(\bar{L}^*) + o(1)$.

For the proof of $S_{L,1}(b) = o_P\left(n^{-4/5}\right)$, uniformly for $b \in I_n^*$, we consider the process $x \to Z_{T,n}(x)$, with

$$Z_{t,n}(x) = n^{4/5} \int_0^t \int_0^t \bar{H}_{L,xn^{-1/5}}(u,v)\ dM(u)dM(v) - n^{4/5} \int_0^t \bar{H}_{L,xn^{-1/5}}(u,u)\ \alpha(u)\gamma(u)\ du,$$

with $x \in [a_1^*, a_2^*]$. Note that $Z_{T,n}(x) = n^{4/5}\ S_{L,1}(xn^{-1/5})$. Thus for for $S_{L,1}(b) = o_P\left(n^{-4/5}\right)$ we have to show that

$$\sup_{x \in [a_1^*, a_2^*]} |Z_{T,n}(x)| \quad = \quad o_P\left(1\right). \tag{10}$$

We first show pointwise convergence

$$Z_{T,n}(x) \quad = \quad o_P\left(1\right) \tag{11}$$

for $x \in [a_1^*, a_2^*]$. Now, for $x \in [a_1^*, a_2^*]$ fixed, we get that $t \to Z_{t,n}(x)$ is a martingale. This follows from the representation: $Z_{t,n}(x) = \int_0^t R_n(w,x)\ dM(w)$ with $R_n(w,x) = n^{4/5} \int_0^w 2\bar{H}_{L,b}(u,w)\mathrm{I}[u \neq w]\gamma(u)\ dM(u) - \bar{H}_{L,b}(w,w)$ and the fact that $R_n(t,x)$ is predictable with respect to the filtration $(\mathcal{F}_t)_{t \geq 0}$. For the proof of (11) we will apply a martingale central limit theorem. We state a central limit theorem instead of a simpler law of large numbers because we will make use of the central limit theorem again below. Suppose that for some $\sigma^2 \geq 0$ and a martingale $V_{t,n} = \int_0^t W_n(w)\ dM(w)$ with an $(\mathcal{F}_t)_{t \geq 0}$ predictable process $W_n(t)$ we have that:

$$\int_0^T W_n^2(t) Y^{(n)}(t) \alpha(t)\ dt = \sigma^2 + o_P(1), \tag{12}$$

$$\int_0^T W_n^2(t) \mathrm{I}[W_n^2(t) > \varepsilon] Y^{(n)}(t) \alpha(t)\ dt = o_P(1) \text{ for all } \varepsilon > 0. \tag{13}$$

Then it holds that

$$V_{T,n} = \int_0^T W_n(w)\ dM(w) \to \mathrm{N}(0, \sigma^2), \text{ in distribution.} \tag{14}$$

For a discussion of this central limit theorem, see e.g. Ramlau-Hansen (1983). In the proof of (11) we apply this result with $\sigma^2 = 0$ and $W_n(t) = R_n(t,x)$. Then (12) implies (13) and for (11) we have to show that $\int_0^T R_n^2(t,x) Y^{(n)}(t) \alpha(t)\ dt = o_P(1)$ for all $x \in [a_1^*, a_2^*]$. Because of (A1) this claim follows from $n \int_0^T R_n^2(t,x)\ dt = o_P(1)$. This immediately follows from $n\ \mathrm{E}[\int_0^T R_n(t,x)^2\ dt] = o(1)$. This concludes the proof of (11).

For the proof of (10) we have to show that the process $x \to Z_{T,n}(x)$ is tight. For this purpose we apply the tightness criterion (12.51) in Billingsley (1968). For (10) it suffices to show that $\mathrm{E}[\{Z_{T,n}(x_1) - Z_{T,n}(x_2)\}^2] \leq C(x_1 - x_2)^2$ for some constant C and for all $x_1, x_2 \in [a_1^*, a_2^*]$. This inequality can be shown by a direct calculation. This concludes the proof of Lemma 3.

For the asymptotic discussion of the cross-validation selector $\widehat{b}_{CV}^{L}$ note that minimizing $\widehat{Q}_L(b)$ is equivalent to the minimization of

$$\widehat{\Delta}_L(b) = \widehat{Q}_L(b) + n^{-1} \int \alpha(t)^2 w(t) Y^{(n)}(t) dt + 2n^{-1} \int \alpha(t) w(t)\ dM(t).$$

The two last terms in $\widehat{\Delta}_L(b)$ do not depend on the bandwidth b. Thus, the minimizer of $\widehat{\Delta}_L(b)$ is equal to the CV-bandwidth $\widehat{b}_{CV}^{L}$. Note also that it holds that

$$\widehat{Q}_L(b) = n^{-1} \left\{ \int_0^T \left[\widehat{\alpha}_{b,L}(s)\right]^2 Y^{(n)}(s) w(s) ds - 2 \int_0^T \widehat{\alpha}_{b,L}^{-}(s) w(s)\ dN(s) \right\},$$

where $\widehat{\alpha}_{b,L}^{-}(t) = \int_0^T \bar{L}_{t,b}(t-s) W(s) \mathrm{I}[s \neq t]\ dN(s)$.

The next lemma states consistency of cross-validation.

Lemma 4. *Under A1–A3, we get for $L = K$ and $L = L_j$ $(j = 1, ..., J)$ that $D_{1,L}(b) = \Delta_L(b) - \widehat{\Delta}_L(b) = o_P\left(n^{-4/5}\right)$, uniformly for $b \in I_n^*$. In particular, we have that $\widehat{b}_{CV}^{L} = b_{MISE}^{L} + o_P\left(n^{-1/5}\right)$.*

Proof. For brevity we write, as in the proof of Lemma 3, $\widehat{\alpha} = \widehat{\alpha}_{b,L}$, $\widehat{\alpha}^{-} = \widehat{\alpha}_{b,L}^{-}$ and $\alpha^* = \alpha_{b,L}^*$. By simple calculations one gets that

$$\begin{aligned} nD_{1,L}(b) &= 2\int_0^T [\widehat{\alpha}^{-}(s) - \alpha(s)] w(s)\ dM(s) + 2\int_0^T [\widehat{\alpha}^{-}(s) - \widehat{\alpha}(s)] w(s) Y^{(n)}(s) \alpha(s)\ ds \\ &= 2n^{-1} \int_0^T [\widehat{\alpha}^{-}(s) - \alpha(s)] w(s)\ dM(s) \\ &= 2\int_0^T [\widehat{\alpha}^{-}(s) - \alpha^*(s)] w(s)\ dM(s) + 2\int_0^T [\alpha^*(s) - \alpha(s)] w(s)\ dM(s) \\ &= nU_{1,L}(b) + nU_{2,L}(b). \end{aligned}$$

We will show that

$$U_{1,L}(b) = U_{1,L}^*(b) + o_P(n^{-4/5}) \tag{15}$$

uniformly for $b \in I_n^*$, where $U_{1,L}^*(b) = 2n^{-1} \int_0^T \widehat{d}^{-}(s) w(s)\ dM(s)$ and $\widehat{d}^{-}(t) = \sum_{i=1}^n \int_0^T \bar{L}_{t,b}^{+}(t-s)\mathrm{I}(s \neq t) W(s)\ dM_i(s)$ with $\bar{L}_{t,b}^{+}$ defined as in the proof of Lemma 3.

With similar arguments as in the proof of Lemma 3 one can show that both terms, $U_{1,L}^*(b)$ and $U_{2,L}(b)$ are of order $o_P(n^{-4/5})$. This gives $D_{1,L}(b) = o_P(n^{-4/5})$. Using similar arguments as in the proof of Lemma 3 one can show that the convergence is uniform. This implies the statement of Lemma 4. It remains to show (15).

For the proof of (15) we apply Lemma A1 in Mammen and Nielsen (2007). For this purpose we write: $U_{1,L}(b) - U_{1,L}^*(b) = \sum_{i=1}^n \int_0^T h_i(s)\ dM_i(s)$ with $h_i(s) = n^{-1} \sum_{j \neq i} \int_0^T (\bar{L}_{s,b} - \bar{L}_{s,b}^{+})(s-t)\ dM_j(t)$. According to Lemma A1 in Mammen and Nielsen (2007) it holds that

$$E(U_{1,L}(b) - U_{1,L}^*(b))^2 \leq \sum_{i=1}^n \rho_i^2 + n \sum_{i=1}^n \delta_i^2,$$

where $\rho_i^2 = E[\int_0^T h_i^2(s)\ Y_i(s)\alpha(s)\ ds]$ and $\delta_i^2 \equiv E[\int_0^T \left(n^{-1}\int_0^T (\bar{L}_{s,b} - \bar{L}_{s,b}^+)(s-t)\ dM_j(t)\right)^2 Y_i(s)\alpha(s)\ ds]$ with $j \neq i$. We now use that because of Assumption (A1) for a constant $C > 0$ we have that $|(\bar{L}_{s,b} - \bar{L}_{s,b}^+)(s-t)| \leq Cn^{-1}n^{-2/5}(\log n)^{-1/2}n^{1/5} = Cn^{-6/5}(\log n)^{-1/2}$. This implies the following bound for δ_i^2 with some constants $C_1, C_2, ... > 0$:

$$\begin{aligned}\delta_i^2 &\leq C_1 n^{-2}n^{-12/5}(\log n)^{-1} \\ &\quad \times E\left[\int_0^T \int_0^T \int_0^T I(|t-s| \leq C_2 n^{-1/5}) I(|u-s| \leq C_2 n^{-1/5})\ dM_j(t)\ dM_j(u) ds\right] \\ &\leq C_3 n^{-22/5}(\log n)^{-1} \int_0^T \int_0^T I(|t-s| \leq C_2 n^{-1/5})\ dt\ ds \leq C_4 n^{-23/5}(\log n)^{-1}.\end{aligned}$$

By using similar bounds one gets that $\rho_i^2 \leq C_5 n^{-18/5}(\log n)^{-1}$. Thus, we have that $E(U_{1,L}(b) - U_{1,L}^*(b))^2 \leq C_6 n^{-13/5}(\log n)^{-1}$.We now argue that for two bandwidths b_1, b_2 with $|b_1 - b_2| \leq n^{-3/5}(\log n)^{-1/2}$ it holds that $|[U_{1,L}(b_1) - U_{1,L}^*(b_1)] - [U_{1,L}(b_2) - U_{1,L}^*(b_2)]| \leq C_7 n^{-2/5}(\log n)^{-1/2}|b_1 - b_2|n^{1/5} \leq C_7 n^{-4/5}(\log n)^{-1}$, where again Assumption (A1) has been used. The last inequality implies that it suffices to show $\sup_{b \in I_n^{**}} |U_{1,L}(b) - U_{1,L}^*(b)| = o_P(n^{-4/5})$, where I_n^{**} is a finite subset of I_n^* with less than $C_8 n^{2/5}(\log n)$ elements. Here, I_n^{**} can be chosen as a grid of points that is contained in I_n^* and where neighbored points have a distance less than or equal to $n^{-3/5}(\log n)^{-1}$. Now for $\delta > 0$ it holds that $P\left(\sup_{b \in I_n^{**}} |U_{1,L}(b) - U_{1,L}^*(b)| > \delta n^{-4/5}\right) \leq \sum_{b \in I_n^{**}} P\left(|U_{1,L}(b) - U_{1,L}^*(b)| > \delta n^{-4/5}\right)$ $\leq \sum_{b \in I_n^{**}} E\left|U_{1,L}(b) - U_{1,L}^*(b)\right|^2 \delta^{-2} n^{8/5} \leq C_8 n^{2/5}\ (\log n) C_6 n^{-13/5}(\log n)^{-1}\delta^{-2}n^{8/5} = C_9$ $\delta^{-2}n^{-3/5}$. Because this upper bound converges to 0 we get that (15) holds. This concludes the proof of the lemma.

The next lemmas enable us to develop linear expansions of $\widehat{b}_{ISE}^L$. For functions G depending on the bandwidth b we denote by G' and G'' the first or second derivative of this function w.r.t. b, respectively.

LEMMA 5. *Under A1–A3, we get for $L = K$ and $L = L_j$ $(j = 1, ..., J)$ that $\Delta_L''(b) = M_L''(b) + o_P\left(n^{-2/5}\right)$ and $D_{1,L}''(b) = o_P\left(n^{-2/5}\right)$ uniformly for $b \in I_n^*$.*

PROOF. This lemma can be shown with similar arguments as used in the proof of Lemma 4. Note first that the derivative of a kernel $R_b(u) = b^{-1}R(b^{-1}u)$ w.r.t. to the bandwidth b is equal to $b^{-2}R^*(b^{-1}u) = b^{-1}R_b^*(u)$ with $R^*(u) = -R(u) - uR'(u)$ and that the second derivative is equal to $b^{-2}R_b^{**}(u)$ with $R^{**}(u) = 2R(u) + 4uR'(u) + u^2R''(u)$. Thus the first and the second derivative behave like the product of a kernel and the factor b^{-1} or b^{-2}, respectively. By looking at the derivatives of $a_{l,b}^*(t)$ and $a_{l,b}^*(t)$ with $l = 0, 1, 2$ one can see that the same holds true for the kernels $\bar{L}_{t,b}$ and $\bar{L}_{t,b}^+$. Using these facts one can easily treat $D_{1,L}''(b)$ as in the proof of Lemma 4. One writes $D_{1,L}''(b)$ as the sum of two expressions and one shows that the integrand in the first expression can be replaced by a predictable integrand, with an error that is now of the order $o_P(n^{-2/5})$, uniformly over all b. The order of the error term is now by a factor $n^{2/5}$ larger because of the just outlined argument. Afterwards one argues again as in the proof of Lemma 3 that the modified first term and the second term are of order $o_P(n^{-2/5})$, uniformly over all b, where the error rate is again increased by a factor $n^{2/5}$ for the same reason as above.

For the expansion of $\Delta_L''(b)$ one replaces $\widehat{\alpha}(t)-\alpha^*(t)$ and its derivatives by $\widehat{d}^*(t)$ and its derivatives. Using brute force bounds as in the proof of Lemma 3 one gets that $\Delta_L''(b) = S_{L,1}''(b)+S_{L,2}''(b)+T_{L,1}''(b)+T_{L,2}''(b)+o_P(n^{-2/5})$, where $S_{L,1}(b)$, $S_{L,2}(b)$, $T_{L,1}(b)$ and $T_{L,2}(b)$ are defined as in Lemma 3.

One now shows that $S_{L,1}''(b) = o_P\left(n^{-2/5}\right)$, $S_{L,2}''(b) = o_P\left(n^{-2/5}\right)$, $T_{L,1}''(b) = 2n^{-1}b^{-3}R(\bar{L}^*)$ $\int_0^T \alpha(t)w(t)dt + o_P\left(n^{-2/5}\right)$, $T_{L,2}''(b) = 3b^2\mu_2^2(\bar{L}^*)\int_0^T \alpha''(t)^2\gamma(t)w(t)dt + o_P\left(n^{-4/5}\right)$. The proof of the first two claims is similar to the proof of (10). Furthermore, the last two statements follow by standard kernel smoothing theory. Note that $M_L''(b) = 3b^2\mu_2^2(\bar{L}^*)\int_0^T \alpha''(t)^2$ $\gamma(t)w(t)dt + 2n^{-1}b^{-3}R(\bar{L}^*)\int_0^T \alpha(t)w(t)dt$. This concludes the proof of Lemma 5.

We now state expansions of $\Delta_L'(b_{MISE}^L)$ and $D_{L,1}'(b_{MISE}^L)$.

Lemma 6. *Under A1-A3, we get for $L = K$ and $L = L_j$ $(j = 1, ..., J)$ that with $b = b_{MISE}^L$*

$$\begin{aligned}
\Delta_L'(b) &= -n^{-2}b^{-2}\int H_L(b^{-1}(u-v))w(u)\gamma(u)^{-1}\, dM(u)\, dM(v) \\
&\quad +2n^{-1}b\mu_2(\bar{L}^*)\int \alpha''(u)w(u)\, dM(u) + o_P\left(n^{-7/10}\right), \\
D_{L,1}'(b) &= -n^{-2}b^{-2}\int G_L(b^{-1}(u-v))w(u)\gamma(u)^{-1}\, dM(u)\, dM(v) \\
&\quad +2n^{-1}b\mu_2(\bar{L}^*)\int \alpha''(u)w(u)\, dM(u) + o_P\left(n^{-7/10}\right),
\end{aligned}$$

where $\bar{L}_b^*(u) = b^{-1}L^*(b^{-1}u)$, $\bar{L}_b^{**}(u) = b^{-1}L^{**}(b^{-1}u)$, $G_L(w) = I[w \neq 0](\bar{L}^{**}(w)-\bar{L}^{**}(-w))$ and $H_L(w) = I[w \neq 0]\int \bar{L}^*(u)(\bar{L}^{**}(w+u)-\bar{L}^{**}(-w+u))\, du$ with $\bar{L}^*(u) = \frac{\mu_2(L)-\mu_1(L)u}{\mu_2(L)-(\mu_1(L))^2}L(u)$ and $\bar{L}^{**}(u) = -\frac{\mu_2(L)-\mu_1(L)u}{\mu_2(L)-(\mu_1(L))^2}(L(u)+uL'(u)) + \frac{\mu_1(L)u}{\mu_2(L)-(\mu_1(L))^2}L(u)$. In particular, it holds that $\Delta_L'(b) = O_P\left(n^{-7/10}\right)$ and $D_{L,1}'(b) = O_P\left(n^{-7/10}\right)$.

Proof. We treat $\Delta_L'(b)$ and $D_{L,1}'(b)$ in two steps. In a first step one can use the results in Mammen and Nielsen (2007) to show that replacing the kernels $\bar{L}_{t,b}(u)$ and $\partial_b\bar{L}_{t,b}(u)$ by the kernels $\bar{L}_{t,b}^+(u)$ or $\partial_b\bar{L}_{t,b}^+(u)$, respectively, leads to an error of order $o_P\left(n^{-7/10}\right)$. The arguments are similar to the proof of Lemma 4. But the argumentation is now simpler because we expand the functions $\Delta_L'(b)$ and $D_{L,1}'(b)$ for one value of b and not uniformly for a set of values of b. In a next step the kernels $\bar{L}_{t,b}(u)$ and $\partial_b\bar{L}_{t,b}(u)$ are replaced by the kernels $(W(t)\gamma(t))^{-1}\bar{L}_b^*(u)$ or $(W(t)\gamma(t)b)^{-1}\ \bar{L}_b^{**}(u)$, respectively. This gives an additional error term of order $o_P\left(n^{-7/10}\right)$. This can be proved by the calculation of the first two moments of the approximations of $\Delta_L'(b)$ and $D_{L,1}'(b)$. Note that the calculation of the fist two moments is simplified by the first step because some non-predictable integrands have been replaced by predictable functions. A check of the last approximation shows the statement of the lemma.

Lemma 7. *Under A1-A3, we get for $L = K$ and $L = L_j$ $(j = 1, ..., J)$ that with $b = b_{MISE}^L$ the following two expansions hold: $\widehat{b}_{ISE}^L = b + C_{1,L}^{-1}n^{-8/5}b^{-2}\int H_L(b^{-1}(u-v))w(u)\gamma(u)^{-1}\, dM(u)\, dM(v) - 2n^{-3/5}C_{1,L}^{-1}b\mu_2(\bar{L}^*)\int \alpha''(u)w(u)\, dM(u) + o_P\left(n^{-3/10}\right)$ and $\widehat{b}_{CV}^L = b + C_{1,L}^{-1}n^{-8/5}b^{-2}\int [H_L - G_L](b^{-1}(u-v))\frac{w(u)}{\gamma(u)}\, dM(u)\, dM(v) + o_P\left(n^{-3/10}\right)$, where $C_{1,L} = 5R(\bar{L}^*)^{2/5}\mu_2^{6/5}(\bar{L}^*)\left\{\int_0^T \alpha(t)w(t)dt\right\}^{2/5}\left\{\int_0^T \alpha''(t)^2\gamma(t)w(t)dt\right\}^{3/5}$.*

PROOF. Lemmas 3 – 6 imply that $\widehat{b}^L_{ISE} = b^L_{MISE} - M''_L(\widehat{b}^{L,*})^{-1}\Delta'_L(b^L_{MISE}) + o_P\left(n^{-3/10}\right)$ and $\widehat{b}^L_{CV} = b^L_{MISE} - M''_L(\widehat{b}^{L,**})^{-1}\widehat{\Delta}'_L(b^L_{MISE}) + o_P\left(n^{-3/10}\right)$, where the bandwidth $\widehat{b}^{L,*}$ lies between $\widehat{b}^L_{ISE}$ and b^L_{MISE} and where the bandwidth $\widehat{b}^{L,**}$ lies between $\widehat{b}^L_{CV}$ and b^L_{MISE}. The claim of the lemma now follows from the expansions of Lemma 6, the continuity of M''_L and using $M''_L(b^L_{MISE}) = C_{1,L}n^{-2/5}$.

PROOF (OF THEOREM 1). Lemma 7 implies that with $b = b^K_{MISE} = \rho_j b^{L_j}_{MISE}$ we get for $\widehat{b}^* = \sum_{j=1}^J \omega_j\rho_j\hat{b}^{L_j}_{CV}$ that $\widehat{b}^* - b = B_1(T) + o_P\left(n^{-3/10}\right)$ and $\widehat{b}^* - \widehat{b}^K_{ISE} = B_2(T) + B_3(T) + o_P\left(n^{-3/10}\right)$, where $B_1(t) = n^{-8/5}b^{-2}\sum_{j=1}^J \omega_j\rho_j^3 C^{-1}_{1,L_j}\int_0^t\int_0^t(H_{L_j} - G_{L_j})(\rho_j b^{-1}(u-v))w(u)\gamma(u)^{-1}\,dM(u)\,dM(v)$, $B_2(t) = n^{-8/5}b^{-2}\int_0^t\int_0^t[\sum_{j=1}^J \omega_j\rho_j^3 C^{-1}_{1,L_j}(H_{L_j} - G_{L_j})(\rho_j b^{-1}(u-v)) - C^{-1}_{1,K}H_K(b^{-1}(u-v))]w(u)\gamma(u)^{-1}\,dM(u)\,dM(v)$ and $B_3(T) = 2n^{-3/5}bC^{-1}_{1,K}\mu_2(K)\int_0^t\alpha''(u)w(u)\,dM(u)$. The statement of the theorem follows by application of a martingale central limit theorem to the martingales $B_1(t)$ and $B_2(t)$, see (12)–(14). We omit the details for checking (12)–(13). It can be shown that $B_1(T)$ and $B_2(T)$ have asymptotic variances σ_1^2 or σ_2^2, respectively.

References

Aalen, O. O. (1978). Non-parametric inference for a family of counting processes. *Ann. Statist.*, **6**, 701–726.

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

Bagkavos, D. (2011) Local linear hazard rate estimation and bandwidth selection. *Ann. Inst. Statist. Math.*, **63**, 1019–1046.

Bagkavos, D. and Patil, P. N. (2008). Local polynomial fitting in failure rate estimation. *IEEE Trans. Reliab.*, **57**, 41–52.

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.

Bolance, C., Guillén, M. and Nielsen, J. P. (2003). Kernel density estimation of actuarial loss functions. *Insur. Math. Econ.*, **32**, 19–36.

Borgan, O. (1984). Maximum likelihood estimation in parametric counting process models with applications to censored failure time data. *Scand. J. Statist.*, **11**, 1–16. Correction: 11, 275.

Buch-Kromann, T., Guillén, M., Nielsen, J. P. and Linton, O. (2011). Multivariate density estimation using dimension reducing information and tail flattening transformations. *Insur. Math. Econ.*, **48**, 99–110.

Buch-Larsen, T., Nielsen, J. P., Guillén, M. and Bolance, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, **39**, 503–518.

Clements, A. E., Hurn, S. and Lindsay, K. A. (2003). Möbius-like mappings and their use in kernel density estimation. *J. Am. Statist. Ass.*, **98**, 993–1000.

Gámiz, M. L., Mammen, E., Martínez-Miranda, M. D. and Nielsen, J. P. (2014). *DOvalidation: Local Linear Hazard Estimation with Do-Validated and Cross-Validated Bandwidths.* R package version 0.1.0.

Gámiz, M. L., Martínez-Miranda, M. D. and Nielsen, J. P. (2013a). Smoothing survival densities in practice. *Comput. Statist. Data Anal.*, **58**, 368–382.

Gámiz, M. L., Janys, L., Martínez-Miranda, M. D. and Nielsen, J. P. (2013b). Bandwidth selection in marker dependent kernel hazard estimation. *Comput. Statist. Data Anal.*, **68**, 155–169.

Gandy, A. and Jensen, U. (2005). Checking a semi-parametric additive risk model. *Lifetime Data Anal.*, **11**, 451–472.

González-Manteiga, W., Cao, R. and Marron, J. S. (1996). Bootstrap Selection of the Smoothing Parameter in Nonparametric Hazard Rate Estimation. *J. Am. Statist. Ass.*, **91**, 1130–1140.

Gram, J. P. (1879). Om Rækkeudvikliner, bestæmt ved hjælp af Mindste Kvadraters Metode. *Copenhagen, A.F. Hæst og Sæn.*

Gram, J. P. (1883). Æber Entwichkelung reeller Funktionen in Reihen mittelst der Methode der Kleinsten Quadrate. *J. Reine Angew. Math.*, **94**, 41–73.

Gustafsson, J., Hagmann, M., Nielsen, J. P. and Scaillet, O. (2009). Local transformation kernel density estimation of loss distributions. *J. Bus. Econ. Statist.*, **27**, 161–175.

Hart, J. D. and Lee, C. L. (2005). Robustness of one-sided cross-validation to autocorrelation. *J. Multiv. Anal.*, **92**, 77–96.

Hart, J. D. and Yi, S. (1998) One-Sided Cross-Validation. *J. Am. Statist. Ass.*, **93**, 620–631.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* **B**, **60**, 271–293.

Jeon, Y. and Kim, J. H. T. (2013). A gamma kernel density estimation for insurance loss data. *Insur. Math. Econ.*, **53**, 569–579.

Jiang, J. and Doksum, K. (2003). On Local Polynomial Estimation of Hazard Rates and Their Derivatives under Random Censoring. *Lecture Notes-Monograph Series*, Festschrift for Constance van Eeden **42**, 463–481.

Loader, C.R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.*, **27**, 415–438.

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P. and Sperlich, S. (2011). Do-validation for kernel density estimation. *J. Am. Statist. Ass.*, **106**, 651–660.

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P. and Sperlich, S. (2014). Further theoretical and practical insight to the do-validated bandwidth selector. *J. Korean Statist. Soc.*, **43**, 355–365.

Mammen, E. and Nielsen, J. P. (2007). A general approach to the predictability issue in survival analysis with applications. *Biometrika*, **94**, 873–892.

Martínez-Miranda, M. D., Nielsen, J. P. and Sperlich, S. (2009). One sided cross-validation for density estimation with an application to operational risk. In: *Operational Risk Towards Basel III: Best Practices and Issues in Modelling. Management and Regulation*, 177–195. New Jersey: Gregoriou (eds.), John Wiley and Sons, G.N.

Müller, H. -G., Wang, J. L. (1994). Hazard Rate Estimation under Random Censoring with Varying Kernels and Bandwidths. *Biometrics*, **50**, 61–76.

Müller, H. -G., Wang, J. L. and Capra, W. B. (1997). From Lifetables to Hazard Rates: The transformation approach. *Biometrika*, **84**, 881–892.

Nielsen, J. P. (1990). Kernel Estimation on Densities and Hazards: A Counting Process Approach. *PhD. Thesis*. University of California, Berkeley, USA.

Nielsen, J. P. and Tanggaard, C. (2001). Boundary and bias correction in kernel hazard estimation. *Scand. J. Statist.*, **28**, 675–698.

Nielsen, J. P., Tanggaard, C. and Jones, M. C. (2009). Local linear density estimation for filtered survival data. *Statistics*, **43**, 167–186.

Patil, P. N. (1993). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plannng Inf.*, **35**, 15–30.

Patil, P. N., Wells, M. T and Marron, J S (1994). Some Heuristics of Kernel Based Estimators of Ratio Functions. *J. Nonparametr. Statist.*, **4**, 203–209.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL: `http://www.R-project.org`.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**, 453–466.

Savchuk, O. Y., Hart, J. D., Sheather S. J. (2008). An empirical study of indirect cross-validation. *IMS Lecture Notes - Festschrift for Tom Hettmansperger*.

Savchuk, O. Y., Hart, J. D., Sheather S. J. (2010). Indirect crossvalidation for density estimation. *J. Am. Statist. Ass.*, **105**, 415–423.

Spierdijk, L. (2008). Nonparametric conditional hazard rate estimation: A local linear approach. *Comput. Statist. Data Anal.*, **52**, 2419–2434.

Spreeuw, J., Nielsen, J. P. and Jarner, S. F. (2013). A visual test of mixed hazard models, *SORT*, **37**, 153–174.

Tutz, G. and Pritscher, L. (1996). Nonparametric estimation of discrete hazard functions, *Lifetime Data Anal.*, **2**, 291–308.

Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformation in density estimation (with comments). *J. Am. Statist. Ass.*, **86**, 343–361.

Wang, J. L. (2005). Smoothing hazard rates. Encyclopedia of Biostatistics. In *Encyclopedia of Biostatistics*, 2nd Edition, Ed. P. Armitage and T. Colton, **7** 4986–4997. Chichester: Jonh Wiley and Sons, Ltd.

Wang, J. L., Müller, H. -G. and Capra, W. B. (1998). Analysis of oldest-old mortality: Lifetables revisited. *Ann. Statist.*, **26**, 126–163.