Bandwidth selection in marker dependent kernel hazard estimation☆

M. Luz Gámiz Pérez[a], Lena Janys[b], María Dolores Martínez Miranda[d], Jens Perch Nielsen[c,1,*]

[a] *University of Granada, Spain*
[b] *University of Mannheim, Germany*
[c] *Cass Business School, City University London, U.K.*
[d] *Cass Business School, City University London, U.K.*

Abstract

Practical estimation procedures for local linear estimation of an unrestricted failure rate when more information is available than just time are developed. This extra information could be a covariate and this covariate could be a time series. Time dependent covariates are sometimes called markers, and failure rates are sometimes called hazards, intensities or mortalities. It is shown through simulations and a practical example that the fully local linear estimation procedure exhibits an excellent practical performance. Two different bandwidth selection procedures are developed. One is an adaptation of classical cross-validation, and the other one is indirect cross-validation. The simulation study concludes that classical cross-validation works well on continuous data while indirect cross-validation performs only marginally better. However, cross-validation breaks down in the practical data application to old-age mortality. Indirect cross-validation is thus shown to be superior when selecting a fully feasible estimation method for marker dependent hazard estimation.

Keywords: Local linear estimation, Bandwidth, Cross-validation, Indirect cross-validation, Aalen's multiplicative model, Survival

1. Introduction

Marker dependent hazard estimation is omnipresent in the mathematical statistical literature. Applications exist in many fields, such as actuarial science, applied statistics, biostatistics, econometrics, engineering and finance. The semiparametric structure considered in Cox (1972) and Andersen and Gill (1982) is widely used in the literature and in practice. Additionally, an enormous amount of semi-parametric dynamic survival models can be found in the literature (see for example Andersen et al. (1993), Fleming and Harrington (1991) and Martinussen and Scheike (2009)). We study the fully unspecified multivariate hazard estimation problem, which has received less attention in the literature than semiparametric hazard models. We work with general filtered survival data, allowing for repeated left truncations and right censoring, as well

☆Supplementary material can be found online

*Corresponding author

[1] Cass Business School, City University London Faculty of Actuarial Science and Insurance 106 Bunhill Row, London EC1Y 8TZ E-mail: Jens.Nielsen.1@city.ac.uk Tef. +44 (0) 20 7040 0909 Fax. +44 (0) 20 7040 8572

as fully general time dependent structures on our markers or covariates. Our starting point is the multivariate local linear estimator of Nielsen (1998). It arises from a local linear minimisation principle around the observed counting process, mimicking the delta function approach developed earlier in one-dimensional density estimation by Jones (1993).It is perhaps surprising that a fully feasible estimation procedure has not yet been published for the multivariate local linear estimator (see Nielsen and Tanggaard (2001) and Bagkavos (2011) for bandwidth selectors in the one-dimensional situation). In this paper we develop the classical cross-validation procedure for the marker dependent hazard estimator and we show that it works well in our finite sample studies. However, cross-validation breaks down in our application based on aggregated data. Indirect cross-validation is known to have a better theoretical and practical performance than cross-validation, and it is known to be more robust when applied to discrete data (see Martínez-Miranda et al. (2009), Savchuk et al. (2010), Mammen et al. (2011) and Gámiz et al. (2013) for the related density case. Consequently, in this paper we develop indirect cross-validation for the local linear estimator, which works well when applied to our aggregated data.

The remainder of the paper is organised as follows. In Section 2 we formulate the estimation problem and present the local linear principle following Nielsen (1998). The estimator is formulated in the general counting process formulation. Direct and indirect cross-validation methods are developed in Section 3. The asymptotic theory necessary to implement indirect cross-validation is provided in Appendix A. Simulation experiments are presented in Section 4 and a real data application to old-age mortality is presented in Section 5. These sections are supplemented by Appendix B, which contains discrete approximations of the estimation strategy in order to work with occurrences and exposures. The explicit algorithms used in the simulation experiments are also described there. Some concluding remarks are given in Section 6.

2. The local linear principle for multivariate kernel hazard estimation

In this section we define the local linear marker dependent hazard estimator. We assume that the data follow Aalen's multiplicative intensity model (see Aalen (1978) and Andersen et al. (1993)), which is defined as follows: Let $Z(t)$ be a d-dimensional time dependent covariate or marker dependent process, and let $\lambda(t)$ be the stochastic hazard for an individual with history $\{Z(s); s \leq t\}$. We examine the following marker dependent hazard model:

$$\lambda(t) = \alpha\{t, Z(t)\}Y(t),$$

where $Y(t)$ is an indicator of survival at time t. Suppose we are observing n individuals and let N_i count observed failures for the ith individual in the time interval, which for simplicity is assumed to be $(0, 1)$, for $i = 1, \ldots, n$. Let $\mathbf{N}^{(n)} = (N_1, \ldots, N_n)$ be a n-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_t$, $t \in (0, 1)$, i.e. one that obeys *les conditions habituelles* (see Andersen et al. (1993, pp.60)). The random intensity process $\lambda^{(n)} = (\lambda_1, \ldots, \lambda_n)$ of $\mathbf{N}^{(n)}$ is then modelled as depending on the d-dimensional marker dependent processes $Z_1(t), \ldots, Z_n(t)$ by

$$\lambda_i^{(n)}(t) = \alpha\{t, Z_i(t)\}Y_i(t), \tag{1}$$

with no restriction on the functional form of $\alpha(\cdot)$. Here Y_i is a predictable process taking values in $\{0,1\}$, indicating (by the value 1) when the ith individual is under risk, for $i = 1,\dots,n$. The marker process $Z_i = (Z_{i1},\dots,Z_{id})$ is a d-dimensional, predictable, *CADLAG* covariate. Let $F_s(z) = \Pr(Z_i(s) \le z | Y_i(s) = 1)$ be the conditional distribution function of the covariate process at time s. Furthermore, let $f_s(z)$ be the corresponding density with respect to the d-dimensional Lebesgue measure. We assume that the marker process is supported on the unit cube and that $E\{Y_i(s)\} = y(s)$, where $y(\cdot)$ is continuous. The marker $Z_i(s)$ is only observed for those s where $Y_i(s) = 1$. Let

$$Z_i^*(s) = \begin{cases} Z_i(s) & \text{when} \quad Y_i(s) = 1 \\ -\infty & \text{when} \quad Y_i(s) = 0 \end{cases}$$

and assume that the stochastic processes $(N_1, Z_1^*, Y_1),\dots,(N_n, Z_n^*, Y_n)$ are i.i.d. for n individuals and $\mathcal{F}_t = \sigma(\mathbf{N}^{(n)}(s), \mathbf{Z}(s), \mathbf{Y}(s); s \le t)$, where $\mathbf{Y} = (Y_1,\dots,Y_n)$ and $\mathbf{Z} = (Z_1,\dots,Z_n)$. Hereafter we simplify the notation by writing $x = (t,z)$ and $W_i(s) = \{s, Z_i(s)\}$, both being vectors with dimension $d+1$ and elements enumerated from 0 to d. Let $\mathcal{K}$ be a $d+1$-dimensional kernel and $\underline{b} = (b_0,\dots,b_d)$ a $d+1$-dimensional bandwidth vector. Let $\mathcal{K}_{\underline{b}}(x-y) = |\underline{b}|^{-1}\mathcal{K}\{(x_0-y_0)/b_0,\dots,(x_d-y_d)/b_d\}$, where $x = (x_0,\dots,x_d)$ and $y = (y_0,\dots,y_d)$ are $(d+1)$-dimensional vectors and $|b| = \prod_{j=0}^d b_j$. We restrict ourselves to the case of multiplicative kernels, that is, $\mathcal{K}(u) = \prod_{j=0}^d K_j(u_j)$, where K_j is a univariate kernel.

The local linear estimator of the hazard rate α is then defined as the solution of the following minimisation problem:

$$\begin{pmatrix} \widehat{\Theta}_0 \\ \widehat{\Theta}_1 \end{pmatrix} = \arg\min_{\Theta_0,\Theta_1} \sum_{i=1}^n \int \left[\Delta N_i(s) - \Theta_0 - \sum_{j=0}^d \Theta_{1j}(x_j - W_{ij}(s)) \right]^2 K_{\underline{b}}(x - W_i(s))Y_i(s)ds. \tag{2}$$

Here we have used the notation $\int g(s)\Delta N_i(s)ds \equiv \int g(s)dN_i(s)$ for any function g. By solving the above problem in Θ_0, the estimator can be written as an intuitive ratio of the smoothed occurrences and smoothed exposures given by:

$$\widehat{\alpha}_{\mathcal{K},\underline{b}}(x) = \frac{\sum_{i=1}^n \int_0^1 \{1 - u^t D(x)^{-1} c_1(x)\} \mathcal{K}_{\underline{b}}(x - W_i(s)) dN_i(s)}{\sum_{i=1}^n \int_0^1 \{1 - u^t D(x)^{-1} c_1(x)\} \mathcal{K}_{\underline{b}}(x - W_i(s)) Y_i(s) ds} := \frac{O_{11}(t,z)}{E_{11}(t,z)}, \tag{3}$$

where $c_1(x) = (c_{10}(x), c_{11}(x),\dots,c_{1d}(x))^t$ (here a^t denotes the transpose of the vector a) with

$$c_0(x) = \sum\nolimits_{i=1}^n \int_0^1 \mathcal{K}_{\underline{b}}\{x - W_i(s)\} Y_i(s) ds,$$

$$c_{1k}(x) = \sum\nolimits_{i=1}^n \int_0^1 \mathcal{K}_{\underline{b}}\{x - W_i(s)\}\{x_j - W_{ij}(s)\} Y_i(s) ds,$$

for $k = 0,1,\dots,d$. Moreover, $D(x) = (d_{jk}(x))_{j,k}$ is the $(d+1)\times(d+1)$ matrix with elements:

$$d_{jk}(x) = \sum\nolimits_{i=1}^n \int_0^1 \mathcal{K}_{\underline{b}}\{x - W_i(s)\}\{x_j - W_{ij}(s)\}\{x_k - W_{ik}(s)\} Y_i(s) ds,$$

for $j, k = 0, 1, \ldots, d$.

While our model and the local linear hazard estimator are identical to those of Nielsen (1998), the above representation of the estimator as a simple fraction of local linear smoothed occurrences divided by local linear smoothed exposures is new. This representation will be quite useful when exploring the underlying hazards in real data applications. Sometimes areas of very low exposure mean that the estimator divides two numbers close to zero. Spurious peaks often arise from pure randomness and these peaks will have a tendency to dominate the visual impression from the graph. In our application to an old-age mortality study based on Swedish data, this is exactly what happens. Simply splitting the hazard estimator into occurrences and exposures as we do above makes it possible to analyse them separately and restrict the visual representation of the hazard to the area of interest.

A generalisation of the recent estimators of Spierdijk (2008) and Kim et al. (2010) to our counting process framework are obtained by the above minimisation principle when it is adjusted to be local linear in the covariates only and local constant in the time direction. We call this estimator the local linear local constant estimator, or the LLLC estimator. In Sections 4 and 5 we study its practical performance and compare it with the fully local linear estimator in (3).

3. Choosing the amount of smoothing

Mammen et al. (2011) and Gámiz et al. (2013) point out that indirect cross-validation seems to be superior to classical cross-validation in practice. The next two subsections introduce classical cross-validation and indirect cross-validation, respectively. Cross-validation works for the local linear estimator as well as for the LLLC estimator, whereas indirect cross-validation only works for the local linear estimator.

3.1. The cross-validation method

In this section the local constant cross-validation method of Nielsen and Linton (1995) is adapted to both the local linear estimator and the LLLC estimator. Let $\widehat{\alpha}_{\mathcal{K},\underline{b}}$ denote any kernel estimator of the hazard rate that depends on a vector of bandwidths $\underline{b} = (b_0, b_1, \ldots, b_d)$ and a multivariate kernel $\mathcal{K}$. The bandwidth selection problem is formulated as follows: Ideally, one would like to choose the smoothing parameter vector as the minimiser of

$$Q_0(\underline{b}) = n^{-1} \sum_{i=1}^{n} \int_0^1 \left[\widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\} - \alpha\{s, Z_i(s)\}\right]^2 Y_i(s) ds, \tag{4}$$

which is equivalent to minimising

$$n^{-1} \left\{ \sum_{i=1}^{n} \int_0^1 \left[\widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\}\right]^2 Y_i(s) ds - 2 \sum_{i=1}^{n} \int_0^1 \widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\} \alpha\{s, Z_i(s)\} Y_i(s) ds \right\}.$$

Only the second of these terms depends on the unknown true α. An estimate of the second term could be given by

$$\sum_{i=1}^{n} \int_0^1 \widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\} Y_i(s) dN_i(s),$$

but this estimator is biased due to the correlation between $\widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\}$ and $dN_i(s)$. This problem can be solved by replacing $\widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\}$ by the leave-one-out version $\widehat{\alpha}^{[i]}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\}$, which is the estimator arising when the dataset is changed by setting the stochastic process N_i equal to 0 for all $s \in (0, 1)$. Then, the cross-validation bandwidth estimate is defined as the minimiser of the following cross-validation score:

$$\widehat{Q}_0(\underline{b}) = n^{-1}\left\{\sum_{i=1}^{n}\int_0^1 \left[\widehat{\alpha}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\}\right]^2 Y_i(s)ds - 2\sum_{i=1}^{n}\int_0^1 \widehat{\alpha}^{[i]}_{\mathcal{K},\underline{b}}\{s, Z_i(s)\}dN_i(s)\right\}. \tag{5}$$

Hereafter this bandwidth estimate will be denoted as $\widehat{\underline{b}}_{CV}$.

3.2. *Indirect cross-validation*

Mammen et al. (2011) introduced do-validation for density estimation as the simple average of the left and right one-sided cross-validation bandwidth estimates of Martínez-Miranda et al. (2009). The motivation for all of these forms of indirect cross-validation is that classical cross-validation tends to work better when the smoothing problem is difficult. Thus, the general indirect cross-validation method starts by formulating a more complex estimation problem to estimate the bandwidth. Then, the resulting bandwidth is rescaled to the original estimation problem (see Martínez-Miranda et al. (2009) and Savchuk et al. (2010) for more details).

The extension of the do-validation method to dimensions higher than one complicates the problem. However, the problem can be simplified by assuming a multiplicative structure for the multidimensional kernel, i.e. $\mathcal{K} = \prod_{j=0}^{d} K_j$. Based on this assumption, a multivariate version of the do-validation method of Mammen et al. (2011)), which works with the local linear multivariate hazard estimator given in (3), is formulated in the following:

For any symmetric one-dimensional kernel K, the left-sided kernel is $K_L(u) = 2K(u)I(-\infty, 0)$ and the right-sided kernel is $K_R(u) = 2K(u)I(0, \infty)$. The multiplicative $(d + 1)$-dimensional kernel can be defined by choosing for each component, $j = 0, \ldots, d$, any of the asymmetric versions, K_L or K_R. Therefore, we can build an (asymmetric) $d+1$-dimensional one-sided kernel such that $\prod_{j=0}^{d} K_j$, where each K_j is K_L or K_R (for $j = 0, \ldots, d$). Let us denote such a kernel by $\mathcal{K}_A$. We now consider the local linear hazard estimator $\widehat{\alpha}_{\mathcal{K}_A,\underline{b}}$ as in (3) but with $\mathcal{K}$ replaced by $\mathcal{K}_A$. Accordingly, we define the one-sided cross-validation score as:

$$OSCV(\underline{b}) = n^{-1}\left\{\sum_{i=1}^{n}\int_0^1 \left[\widehat{\alpha}_{\mathcal{K}_A,\underline{b}}\{s, Z_i(s)\}\right]^2 Y_i(s)ds - 2\sum_{i=1}^{n}\int_0^1 \widehat{\alpha}^{[i]}_{\mathcal{K}_A,\underline{b}}\{s, Z_i(s)\}dN_i(s)\right\}.$$

Let $\widehat{\underline{b}}_A$ denote the minimiser of the above score. Note that such a bandwidth vector is not a suitable bandwidth estimate for the original estimator $\widehat{\alpha}_{\mathcal{K},\underline{b}}$. However, such a suitable bandwidth can be derived by rescaling $\widehat{\underline{b}}_A$. The rescaling constant is defined as the ratio of the MISE-optimal bandwidth for $\widehat{\alpha}_{\mathcal{K},\underline{b}}$ and its one-sided version $\widehat{\alpha}_{\mathcal{K}_A,\underline{b}}$. Let this ratio be denoted by C. As we prove in Appendix A, this rescaling ensures that $C\widehat{\underline{b}}_A$ is a consistent estimate for the optimal bandwidth of the estimator $\widehat{\alpha}_{\mathcal{K},\underline{b}}$. Accordingly, the one-sided cross-validation bandwidth is defined as

$$\widehat{\underline{b}}_{A,OSCV} = C\widehat{\underline{b}}_A. \tag{6}$$

In the one-dimensional setting, do-validation averages one-sided cross-validated bandwidths. In our multivariate setting we have 2^{d+1} different one-sided versions of (6). The generalisation of the do-validated bandwidth to our setting is defined as the average of these 2^{d+1} one-sided bandwidth estimates.

Finally, to ensure that we are defining a bandwidth estimator that is feasible in practice, the rescaling constant C should be a known value. We can see that this requirement is satisfied when we consider the local linear estimator, which exhibits a suitable convergence rate. Using the asymptotic theory developed in Nielsen (1998) we can derive the following exact expression for this constant:

$$C = \left(\frac{\kappa_2 / \kappa_1^2}{\bar{\kappa}_2 / \bar{\kappa}_1^2} \right)^{1/(d+5)},$$

where $\kappa_1 = \int u^2 K(u) du$, $\kappa_2 = \int K(u)^2 du$ and $\bar{\kappa}_1$, $\bar{\kappa}_2$ are defined analogously but involving one-sided kernels (K_L or K_R). The details about this calculation are provided in Appendix A. From the above expression we can confirm that the constant C can be derived without any prior information, since it only depends on the chosen and known kernel K.

For example, the rescaling constant for the kernel $K(u) = 3003/2048(1 - u^2)^6 I(-1 < u < 1)$ and a one-dimensional covariate ($d = 1$) becomes $C = 0.5105$. Similarly, for the Epanechnikov kernel $K(u) = 3/4(1 - u^2)I(-1 < u < 1)$ and again $d = 1$ we have the value $C = 0.5232$.

4. Simulation experiments

The following finite sample study shows that the above defined cross-validation method and the particular case of indirect cross-validation (named as the do-validation method) work well when continuous data is available.

4.1. The objectives

The objectives we pursue through simulation experiments are twofold: (1) to compare the local linear (LL) estimator with the local linear and local constant (LLLC) estimator that generalises the estimator by Spierdijk (2008), and (2) to evaluate the performance of the two bandwidth selectors described in Section 3, namely the cross-validation bandwidth $\widehat{\underline{b}}_{CV}$ and the do-validation bandwidth $\widehat{\underline{b}}_{DO}$.

To achieve the first objective we provide a numerical comparison in two scenarios. The first scenario is the ideal one, where we work in the best possible situation for the estimators.

This amounts to finding the best possible bandwidth, $\widetilde{\underline{b}}_0$, for each simulated set of data, in the sense of having the smallest error (4). We also consider a second infeasible strategy for bandwidth selection, the *average best bandwidth* strategy, which amounts to finding the bandwidth $\underline{b}_0$ which minimises the averaged error over S simulated samples. Clearly, such optimal bandwidths will be infeasible in practice, and thus we call the derived estimators infeasible estimators. The second scenario establishes a comparison in a practical situation, where the bandwidth is estimated from the data. Here we consider three fully-feasible estimators: the LLLC estimator with cross-validated bandwidth and the LL estimator with both cross-validated and do-validated bandwidths. Note that do-validation can only be developed for the fully local linear estimator.

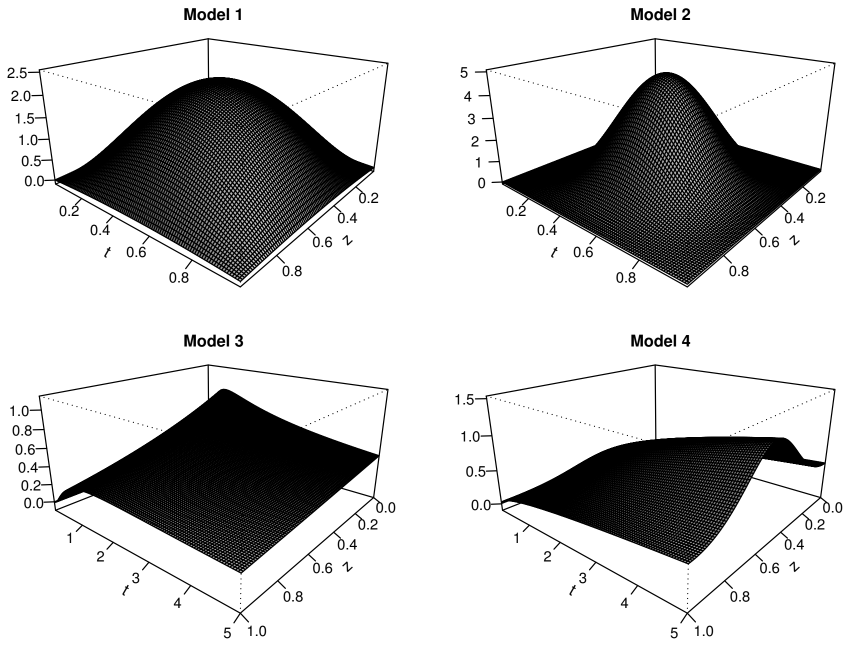

Figure 1: The true two-dimensional hazards in the simulation study.

The second objective is addressed by comparing the two proposed bandwidth estimates for the local linear estimator with the infeasible bandwidth selectors considered as benchmarks. Such a comparison will give some guidance as to how well the feasible bandwidth selectors are doing.

4.2. Simulation settings

We consider a model with a one-dimensional constant marker Z that takes values in the interval $[0, \tau]$. We consider four two-dimensional hazards. The first two were also considered in Nielsen (1998) and the other two were also considered in Spierdijk (2008). The assumed true hazard rates are:

$$\begin{array}{lcll} \alpha_1(t,z) & = & B(t,2,2) \times B(z,2,2), & t, z \in [0,1] \\ \alpha_2(t,z) & = & B(t,4,4) \times B(z,4,4), & t, z \in [0,1] \\ \alpha_3(t,z) & = & \dfrac{1}{t}\dfrac{\psi(\log(t)-z)}{\Phi(z-\log(t))}, & t \in (0,5], z \in [0,1] \\ \alpha_4(t,z) & = & \frac{3}{2}t^{1/2}\exp\left(-\frac{1}{2}\cos(2\pi z)-\frac{3}{2}\right), & t \in (0,5], z \in [0,1] \end{array} \qquad (7)$$

where $B(\cdot, a, a)$ denotes the density of a Beta with shape and scale parameters equal to a, and $\psi(\cdot)$ and $\Phi(\cdot)$ denote the density and the cumulative distribution function of a Standard Normal, respectively. These four true hazards are shown in Figure 1.

We consider two different sampling schemes to provide complete and also filtered samples. The filtration includes right censoring and left truncation. From each model and sampling scheme, we simulate S = 100 samples of size $n = 100, 500, 1000$ and 5000. The explicit algorithms to simulate both types of samples are provided in Appendix B, specifically in subsections Appendix B.2 and Appendix B.3. From these algorithms the data are recorded in a discrete way so that the estimators (local linear and LLLC) can be calculated from the discrete versions provided in subsection Appendix B.1. This simulation strategy is similar to the one considered in Nielsen (1998) and Nielsen and Tanggaard (2001). We discretise the simulated data in a two-dimensional grid into $[0, \tau] \times [0, 1]$ with size $M \times M'$ and $\tau = 1$ for models 1 and 2 and $\tau = 5$ for models 3 and 4. To ensure the stability of the results and therefore the validity of the approximation we choose the values $M = M' = 100$. Bigger grid sizes do not appear to alter our conclusions.

Finally, the kernel estimators are always calculated using the kernel $K(x) = 3003/2048(1 - x^2)^6 I(-1 < x < 1)$. The bandwidth selectors are calculated as minimisers of the corresponding scores into an equally spaced grid of bandwidths. We define a two-dimensional grid of 100 bandwidths, $\underline{b} = (b_0, b_1)$, with $b_0 \in [\tau/n, \tau/2]$ and $b_1 \in [1/n, 0.5]$ for each model.

The comparison among hazard estimators is made by evaluating the following performance measure:

$$err(\widehat{\alpha}_{K,\underline{b}}) = n^{-1} \sum_{r=1}^{M} \sum_{r'=1}^{M'} \left[\widehat{\alpha}_{K,b}(x_{r,r'}) - \alpha(x_{r,r'})\right]^2 E_{r,r'}, \qquad (8)$$

for any estimator $\widehat{\alpha}_{K,\underline{b}}$ with bandwidth $\underline{b} = (b_0, b_1)$ and kernel K. This is just the discrete version of the global measure of the estimation error in (4). From each simulated sample, each estimator provides a value of the error (8), and the comparisons are made from the averaged errors over the S simulated samples in each scenario.

4.3. Simulation Results

In this section we go through the obtained results from the simulation experiments under the scenarios defined above and draw some conclusions about the two objectives we pursue.

Each true hazard is estimated using two types of estimators: the local linear (LL) and the LLLC. The comparison between these two estimators is done in two steps. Firstly, by considering the best possible situation, i.e. at their respective best optimal bandwidths (defined and denoted above by $\widetilde{\underline{b}}_0$ and $\underline{b}_0$). Secondly, we compare the estimators using data-driven bandwidth selection, which would be the case in practice.

The simulation results are shown in Table 1 and Table 2. The numbers in these tables correspond to the averaged (along the S simulated samples) values of the performance measure in (8). The last three columns in the tables correspond to the practical estimators, specifically the LLLC estimator with cross-validated bandwidth ($\widehat{\underline{b}}_{CV}$) and the local linear estimator with do-validated bandwidth ($\widehat{\underline{b}}_{DO}$).

We can clearly see that the local linear estimator beats the LLLC estimator in all cases. Columns 3 and 6 show the percentage of error reductions by the local linear estimator in the best possible cases. Column 9 shows the reduction in the practical situation where the bandwidth is chosen from the data. We can see reductions in the error of up to 40%. This result is not surprising due to the lack of boundary adjustment in the time direction by the LLLC estimator. This issue will also be observed in the data analyses below. When considering the simulated scenarios in more detail, it turns out that the local linear estimator indeed works much better at the boundaries, as can be seen in Figure 2. In this figure we plot the ratio between the two

		Bandwidth choice								
		Optimal (sample)			Optimal (global)			Data-driven		
Model	n	LLLC	LL	Rat.	LLLC	LL	Rat.	LLLC	LL	Rat.
1	100	0.0760	0.0756	0.99	0.0774	0.0760	0.99	0.1006	0.1038	1.03
	200	0.0597	0.0523	0.88	0.0605	0.0530	0.88	0.0650	0.0654	1.01
	500	0.0292	0.0254	0.87	0.0299	0.0257	0.87	0.0381	0.0268	0.71
2	100	0.1032	0.0942	0.91	0.1067	0.0983	0.91	0.2321	0.1179	0.51
	200	0.0687	0.0573	0.84	0.0692	0.0597	0.84	0.1124	0.0663	0.59
	500	0.0368	0.0311	0.85	0.0380	0.0319	0.85	0.0398	0.0360	0.90
3	100	0.0443	0.0369	0.83	0.0450	0.0398	0.83	0.0619	0.0425	0.69
	200	0.0340	0.0247	0.73	0.0342	0.0247	0.73	0.0380	0.0299	0.79
	500	0.0231	0.0139	0.60	0.0231	0.0139	0.60	0.0261	0.0203	0.78
4	100	0.0400	0.0263	0.66	0.0413	0.0274	0.66	0.0510	0.0380	0.74
	200	0.0244	0.0158	0.65	0.0250	0.0162	0.65	0.0298	0.0262	0.88
	500	0.0172	0.0117	0.68	0.0175	0.0118	0.68	0.0241	0.0155	0.65

Table 1: Comparison between the local linear local constant (LLLC) estimator and the local linear (LL) estimator -case of complete samples. The numbers in the table show the average (along the simulated samples) of the performance measure in (8). Columns 5, 8 and 11 show the ratio between the LL and LLLC errors. Columns 3–5 correspond to the estimators calculated at their best optimal bandwidth at each sample ($\underline{\widetilde{b}}_0$). Columns 6–8 correspond to the best global optimal bandwidths ($\underline{b}_0$). Columns 9–11 correspond to the data-driven bandwidth choice, which is cross-validation for LLLC and do-validation for LL.

		Bandwidth choice								
		Optimal (sample)			Optimal (global)			Data-driven		
Model	n	LLLC	LL	Rat.	LLLC	LL	Rat.	LLLC	LL	Rat.
1	100	0.0947	0.0870	0.92	0.0965	0.0902	0.92	0.1641	0.0980	0.60
	200	0.0569	0.0499	0.88	0.0599	0.0534	0.88	0.0830	0.0578	0.70
	500	0.0262	0.0212	0.81	0.0267	0.0216	0.81	0.0478	0.0229	0.48
2	100	0.1280	0.1151	0.90	0.1326	0.1233	0.90	0.2026	0.1530	0.75
	200	0.1007	0.0961	0.95	0.1022	0.0997	0.95	0.1262	0.1047	0.83
	500	0.0521	0.0488	0.94	0.0530	0.0497	0.94	0.1016	0.0523	0.52
3	100	0.0441	0.0362	0.82	0.0456	0.0374	0.82	0.0538	0.0406	0.75
	200	0.0288	0.0193	0.67	0.0289	0.0200	0.67	0.0377	0.0239	0.63
	500	0.0181	0.0102	0.56	0.0181	0.0103	0.56	0.0198	0.0147	0.74
4	100	0.0365	0.0311	0.85	0.0380	0.0330	0.85	0.0443	0.0372	0.84
	200	0.0304	0.0237	0.78	0.0308	0.0254	0.78	0.0357	0.0283	0.79
	500	0.0176	0.0137	0.78	0.0180	0.0138	0.78	0.0245	0.0178	0.73

Table 2: Comparison between the local linear local constant (LLLC) estimator and the local linear (LL) estimator -case of filtered samples (20% censoring and 25% truncation). The reported numbers are defined as in Table 1.

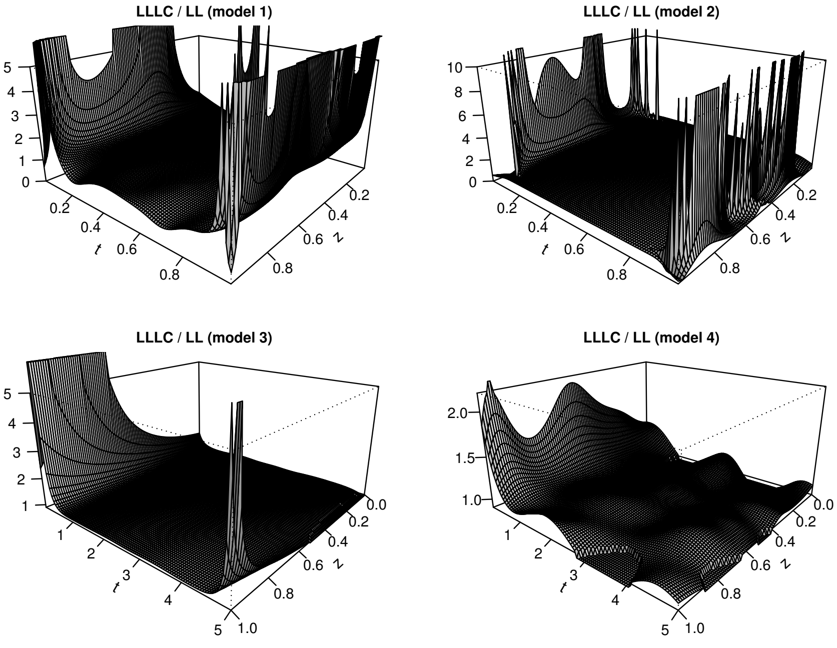

Figure 2: Ratio between the LLLC estimator and the local linear estimator from the 5% best sample for the LLLC estimator (sample size $n = 500$ and no filtration). The estimators are calculated with their respective best optimal bandwidths.

estimators from one simulated sample (the estimators were calculated using their best possible bandwidths with this sample). The sample we choose for the plots is one of the best for the LLLC estimator. Specifically, we choose the sample corresponding to the 5% quantile of the performance measure, which is calculated by ordering the samples according to the reported error by the estimator, and then choosing the sample that gives the highest error of the 5% simulated samples with the best performance.

As a next step we asses the performance of the bandwidth selectors considered for the local linear estimator (our second objective). Table 3 shows the results of the performance measure in (8) for each method in each simulated scenario. To gain a clearer insight into the performance of the different methods, we calculate the relative error with respect to the infeasible best possible bandwidth at each sample ($\widetilde{\underline{b}}_0$). This measure is defined as

$$\mathtt{Rerr} = \frac{averr_{DO} - averr_0}{averr_{CV} - averr_0}, \tag{9}$$

where $averr_\bullet$ is the average of the error measure in (8) along the S simulated samples and considering a specific criterion for the bandwidth selection (CV, DO or $\widetilde{\underline{b}}_0$). The resulting values for each model and the corresponding sample size are reported in columns 5 and 8 of Table 3, for complete samples and filtered samples respectively. Note that Rerr indicates the reduction (if lower than 1) or increase (if bigger than 1) of the error using do-validation instead of standard cross-validation. The conclusion from the results

Table 3: Comparison between cross-validation and do-validation for the local linear estimator. The numbers in the table consist of the average (along the simulated samples) of the performance measure in (8). Columns 5 and 8 show the relative error defined in (9).

		Complete samples			Filtered samples		
Model	n	CV	DO	Rerr	CV	DO	Rerr
1	100	0.1036	0.1038	1.01	0.1887	0.0980	0.11
	200	0.0573	0.0654	2.61	0.0748	0.0578	0.32
	500	0.0385	0.0268	0.11	0.0441	0.0229	0.07
2	100	0.2286	0.1179	0.18	0.1762	0.1530	0.62
	200	0.1008	0.0663	0.21	0.1276	0.1047	0.27
	500	0.0365	0.0360	0.90	0.1052	0.0523	0.06
3	100	0.0553	0.0425	0.31	0.0479	0.0406	0.38
	200	0.0284	0.0299	1.39	0.0269	0.0239	0.61
	500	0.0166	0.0203	2.40	0.0129	0.0147	1.69
4	100	0.0357	0.0380	1.24	0.0342	0.0372	1.93
	200	0.0195	0.0262	2.78	0.0279	0.0283	1.10
	500	0.0197	0.0155	0.48	0.0198	0.0178	0.68

reported in Table 3 is that do-validation outperforms cross-validation most of the time for both complete and filtered samples.

5. An application to old-age mortality

In this section we provide an illustration of the methods described in this paper to the dataset used by Fledelius et al. (2004), which consists of old-age mortality data of Swedish women. Fledelius et al. (2004) argue that local linear estimation is particularly important in studies of old age mortality because of its superior bias properties (compared to local constant estimation). The dataset covers the calendar years 1988 to 1997 and the ages 90 and above. The aim is to estimate the two-dimensional hazard or force of mortality, $\alpha(t,z)$. In our formulation of the hazard function, age is time and calendar year is a one-dimensional covariate. In the study by Fledelius et al. (2004), the estimation was calculated using the same local linear estimator we defined in (3), but with a subjective bandwidth choice. Based on experience, the authors considered a bandwidth parameter for the time dimension of $b_0 = 7$ and $b_1 = 4$ for the covariate. Here we consider two practical data-driven bandwidth choices, namely the cross-validated and the do-validated bandwidths, which were defined in Section 3. To calculate the estimators, we use the Epanechnikov kernel as the symmetric kernel K, which was also used by Fledelius et al. (2004).

Figure 3 shows the resulting smooth two-dimensional hazard estimates of the force of mortality for ages 90 and above between 1988 and 1997. The top left panel shows the local linear estimator with cross-validated bandwidths, which in this case are $\widehat{\underline{b}}_{CV} = (5, 2)$. The hazard estimate using the do-validated bandwidths of $\widehat{\underline{b}}_{DO} = (3.75, 17.25)$ is shown in the top right panel. The LLLC estimator is shown below using the cross-validated bandwidths of $\widehat{\underline{b}}_{CV} = (1, 8)$. From the surfaces in Figure 3, one can see that all methods provide surfaces which are increasing with respect to age. However, they exhibit different shapes and rates of increase. The ratio between the LLLC estimator and the LL estimator is plotted in the right bottom panel in Figure 3. As we would expect, the major differences between the estimators arise at the boundary points. We should also point out that cross-validation (for both the local linear and the LLLC estimators) does not work well in this case. The cross-validation score is minimised at unacceptable undersmoothing levels which lead to very noisy estimates.

In order to gain a better insight into the differences among our estimates, we have plotted in Figure 4 the estimated force of mortality for three different years, 1989, 1992, and 1996.

Each panel in Figure 4 shows the three practical estimates that we are comparing, namely the cross-validated and do-validated local linear estimator and the cross-validated LLLC estimator. We have also plotted the local linear estimator with the bandwidth vector used by Fledelius et al. (2004). The mortality curves in this figure, as well as the surfaces in Figure 3, show that there has been a slight decline in female mortality over the decade which is in line with the results of previous studies. However, the time trend exhibited by each estimate is quite different for the ages 105 and above. Clearly, the LLLC estimator fails dramatically in this age range, although the cross-validated local linear estimator in years 1989 and 1992 seems to perform equally bad.

As discussed above, the fact that the estimator is a ratio (see equation 3) leads to problems in areas with very small exposures. The visualisation technique provided in Figure 5 makes it possible to understand when dramatic peaks might be due to the level of exposures approaching zero. In this figure, the top graphs show the smoothed local linear occurrences and exposures (given by $O_{11}(\cdot,\cdot)$ and $E_{11}(\cdot,\cdot)$ in equation (3)). From this we can see that the estimated mortality above 105 years of age is based on almost no information and the visual analysis becomes difficult. Therefore, the estimators should be compared in the area with more exposure, which we do in the lower graphs of Figure 5. This analysis confirms that the do-validated local linear estimator gives a better estimate of the mortality rates than the estimators based on cross-validated bandwidths. In fact, the do-validated LL estimator provides conclusions similar to the ones guided by expert opinion in Fledelius et al. (2004)).

6. Conclusion

In this paper, a practical and fully feasible methodology for fully nonparametric multivariate local linear hazard estimation has been developed. Such a method did not exist before. First, classical cross-validation was adapted to this setting, and then indirect cross-validation was introduced. Both of these two new feasible methods were shown to work well in finite sample studies with continuous data, but indirect cross-validation was shown to be superior in the practical study of old-age mortality. The difference between the local linear estimator by Nielsen (1998) and the LLLC estimator (which generalises the estimators of Spierdijk (2008) and Kim et al. (2010)) is that the latter is local constant in the time direction. This simplification can have dramatic effects around the boundary regions, where the fully local linear estimator is shown to work much better than the LLLC estimator. This result naturally leads to the research question of whether further bias reduction could be useful around the boundaries. Variable bandwidth methods (see Nielsen (2003)

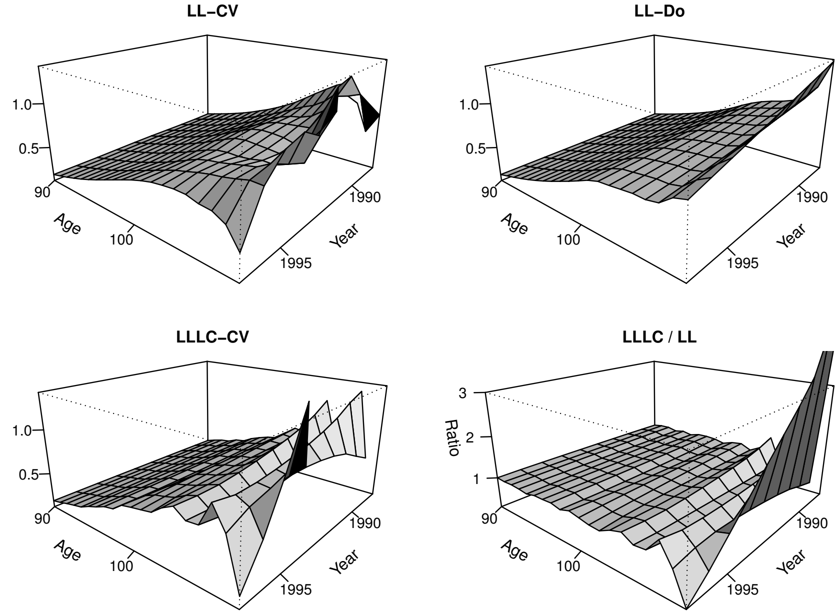

Figure 3: Two-dimensional hazard estimates for the ages 90 and above, between 1988 and 1997. The top left and right panels show the local linear estimator with the cross-validation and the do-validation bandwidths, respectively. The LLLC estimator with cross-validated bandwidth is shown in the right bottom panel. The ratio between the local linear and the LLLC estimators (using their respective cross-validated bandwidths) is shown in the last panel.

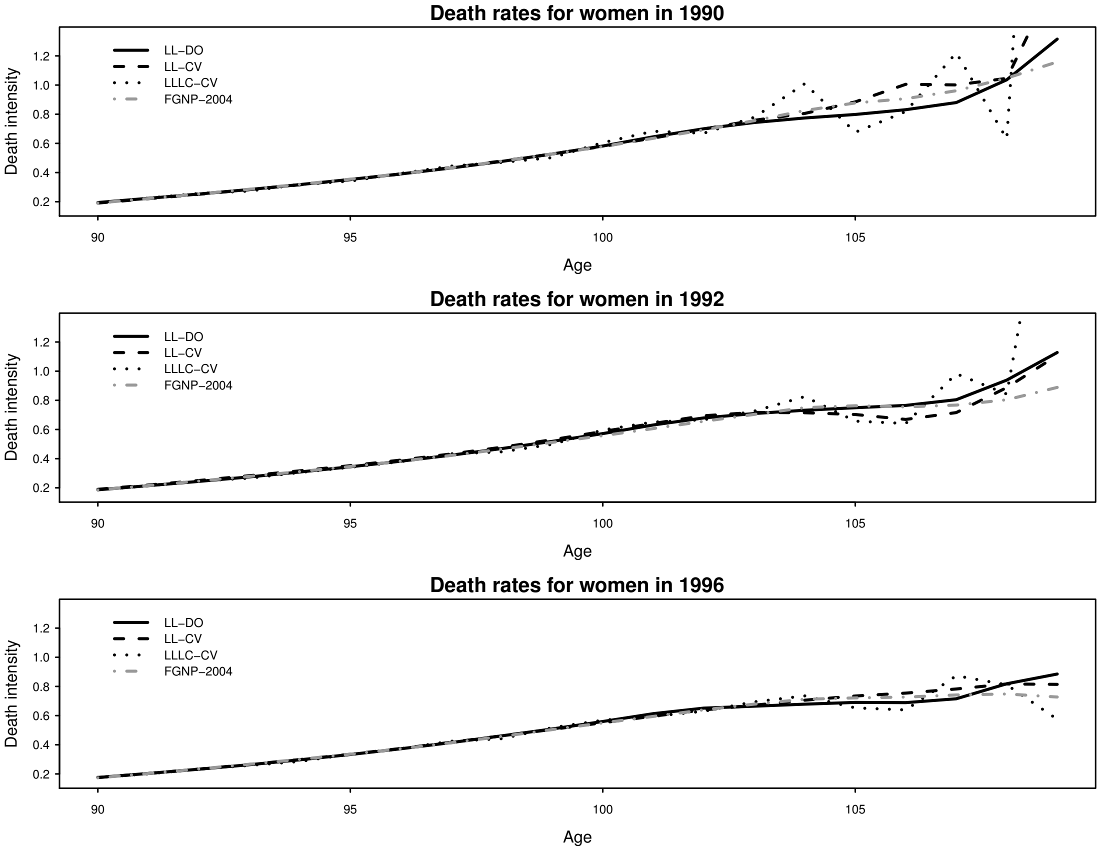

Figure 4: Estimated force of mortality for women age 90-110 for the years 1989, 1992 and 1996.

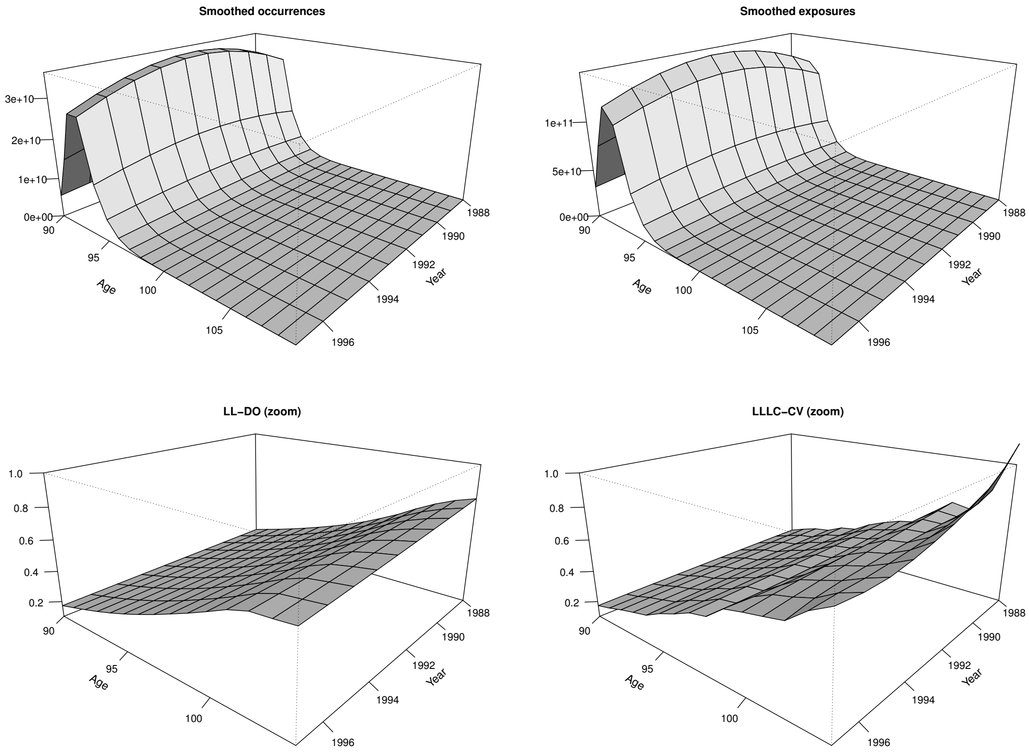

Figure 5: *Guided hazard visualisation by the graphical analysis of occurrences/exposures* (mortality data). The two top graphs show the smoothed exposures and occurrences using local linear estimation and do-validated bandwidths. The bottom graphs show the LLLC estimator and the do-validated local linear estimator, which are now compared in the area where the exposure is significant (women with ages up to 105 years).

and Bagkavos and Patil (2008), transformation methods (see Spreeuw et al. (2013) or multiplicative bias correction methods (see Mammen and Nielsen (2007)) could be useful starting points for further research. It is perhaps surprising that so little focus has been given to provide plots of the fully unrestricted hazard estimator given the enormous popularity of structured hazard models. While structured models, like the popular Cox model, might summarise data in a convenient way, it is also true that important details of the overall hazard might be missed when imposing structure. Therefore, the unrestricted hazard considered in this paper should be recommended as an important part of the applied statisticians tool box when working with survival data.

Acknowledgements

The authors thank the Centro de Servicios de Informática y Redes de Comunicaciones (CSIRC), Universidad de Granada, for providing the computing time (all the calculations have been done in R –R Development Core Team, 2011). We also thank Prof. Guillén (University of Barcelona, Spain) for providing us with the mortality dataset. This work has been financially supported by the Spanish "Ministerio de Ciencia e Innovación" by the grant MTM2008-03010,"Junta de Andalucía" grant TIC-06902 and the European Commission under the Marie Curie Intra-European Fellowship FP7-PEOPLE-2011-IEF Project number 302600. Finally the authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

References

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

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

Andersen, P.K. and Gill, R.D. (1982), Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, 10(4), 1100–1120.

Bagkavos, D. (2011), Local linear hazard rate estimation and bandwidth selection. *Annals of the Institute of Statistical Mathematics*, 63, 1019–1046.

Bagkavos, D. and Patil, P.N. (2008), Local polynomial fitting in failure rate estimation. *IEEE Transactions on Reliability*, 57 (1),41–52.

Buch-Kromann, T. and Nielsen, J.P. (2012), Multivariate density estimation using dimension reducing information and tail flattening transformations for truncated or censored data. *Annals of the Institute of Statistical Mathematics*, 64(1), 167–192.

Cox, D.R. (1972), Regression Models and Life Tables (with Discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Fledelius, P., Guillén, M., Nielsen, J.P. and Petersen, K.S. (2004), A Comparative Study of Parametric and Nonparametric Estimators of Old-Age Mortality in Sweden. *Journal of Actuarial Practice* 2004(1), 101–126.

Fleming, T.R. and Harrington, D.P. (1991), *Counting Processes and Survival Analysis*. Wiley, New York.

Gámiz, M.L., Martínez-Miranda, M.D. and Nielsen, J.P. (2013), Smoothing survival densities in practice. *Computational Statistics and Data Analysis*, 58, 368–382.

Jones, C. (1993) Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3, 135–146.

Kim, C., Oh, M., Yang, S.J. and Choi, H. (2010), A local linear estimation of conditional hazard function in censored data. *Journal of the Korean Statistical Society*, 39, 347–355.

Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P. and Sperlich, S. (2011), Do-validation for kernel density estimation. *Journal of the American Statistical Association*, 106, 651–660.

Mammen, E. and Nielsen, J.P. (2007), A general approach to the predictability issue in survival analysis with applications, *Biometrika*, 94(4), 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: G.N. Gregoriou (eds.) *Operational Risk Towards Basel III: Best Practices and Issues in Modelling. Management and Regulation*, 177–195. John Wiley and Sons, New Jersey.

Martinussen, T. and Scheike, T.H. (2009), The additive hazards model with high-dimensional regressors. *Lifetime Data Analysis*, 15, 330–342.

Nielsen, J.P. (2003), Variable bandwidth kernel hazard estimators, *Journal of Nonparametric Statistics* 3, 355–376.

Nielsen, J.P. and Linton, O.(1995), Kernel estimation in a non-parametric marker dependent hazard model. *Annals of Statistics*, 23, 1735–1748.

Nielsen, J.P. (1998), Marker dependent kernel estimation from local linear estimation. *Scandinavian Actuarial Journal*, 2, 113–224.

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

R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL: http://www.R-project.org

Sánchez-Sellero, C., González-Manteiga, W. and Cao, R. (1999), Bandwidth selection for kernel density estimation with truncated and censored data. *Annals of the Institute of Statistical Mathematics*, 51(1), 51–70.

Savchuk, O.Y., Hart, J.D. and Sheather, S. (2010), Indirect cross-validation for Density Estimation. *Journal of the American Statistical Association*, 105, 415–423.

Spierdijk, L. (2008), Nonparametric conditional hazard rate estimation: A local linear approach. *Computational Statistics and Data Analysis,* 52, 2419–2434.

Spreeuw, J., Nielsen, J.P. and Jarner, S.F. (2013) A visual test of mixed hazard models, *SORT-Statistics and Operations Research Transactions*, in press.

Appendix A. Asymptotic properties of the local linear hazard estimator

The asymptotic properties of $\widehat{\alpha}_{\mathcal{K},\underline{b}}$ are derived in Nielsen (1998) and are established as follows: Let $\varphi(x) = f_t(z)y(t)$ and $\varphi(x) > 0$, additionally we assume the following regularity conditions:

(i) The hazard function, α and φ are fourth and twice continuously differentiable at x, respectively.
(ii) The kernel $\mathcal{K}$ has bounded support, is symmetric around zero and is continuous.
(iii) Suppose that $n|\underline{b}| \to \infty$ and $\underline{b} \to 0$, where $|\underline{b}| = \prod_{j=0}^{d} b_j$.
(iv) There exist a function $\gamma \in C_1([0,\tau])$ positive in t which is the limit of the exposure function that is $\sup_{s\in[0,\tau]} \left|Y^{(n)}(s)/n - \gamma(s)\right| \xrightarrow{P} 0$. Moreover $\sup_{s\in[0,\tau]} \left|nE\left\{I(Y^{(n)}(s))/Y^{(n)}(s)\right\}^{-1} - \{\gamma(s)\}^{-1}\right| \longrightarrow 0$.

Let the kernel moments be defined as: $\kappa_{1j} = \int_{R^{d+1}} v_j^2 \mathcal{K}(v)dv$, for $j = 0, \ldots, d$, and $\kappa_2 = \int_{R^{d+1}} \mathcal{K}(v)^2 dv$. Then the pointwise asymptotic distribution of $\widehat{\alpha}_{\mathcal{K},\underline{b}}(x)$ is analysed by splitting the error $\widehat{\alpha}_{\mathcal{K},\underline{b}}(x) - \alpha(x)$ into the variable part and a stable part, which leads to the following

$$(n|\underline{b}|)^{1/2}\{\widehat{\alpha}_{\mathcal{K},\underline{b}}(x) - \alpha(x) - B_x\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_x^2)$$

where

$$B_x = \sum_{j=0}^{d} \kappa_{1j} b_j^2 \left\{\frac{1}{2}\frac{\partial^2 \alpha(x)}{\partial x_j^2}\right\} + \mathrm{o}\left(\sum_{j=0}^{d} b_j^2\right),$$

and

$$\sigma_x^2 = \kappa_2 \frac{\alpha(x)}{\varphi(x)}$$

The required symmetry for the kernel $\mathcal{K}$ in assumption (ii) is not essential in developing these asymptotics; however, it allows for simpler expressions. In fact, in the limit the stochastic local linear kernel $\mathcal{K}_{x,\underline{b}}(u)$ simply becomes the deterministic $\mathcal{K}(u)/\varphi(x)$. For asymmetric kernels these asymptotic results hold in exactly the same way apart from the involved kernel constants κ_2 and κ_{1j}, $j = 0, 1, \ldots, d$. Specifically, these constants take on the values $\bar{\kappa}_2 = \int [\mathcal{K}^*(u)]^2 \, du$ and $\bar{\kappa}_{1j} = \int u_j^2 \mathcal{K}^*(u)du$, by involving a deterministic equivalent kernel $\mathcal{K}^*(u)$ which is rather more complicated. For the particular case that $d = 1$ this equivalent kernel can be written as

$$\mathcal{K}^*(u_0, u_1) = \frac{1}{\varphi(x)} \frac{\mu_2(K) + \mu_1(K)^2 - \mu_1(K)(u_0 + u_1)}{\mu_2(K) - \mu_1(K)^2} K(u_0)K(u_1),$$

where $\mu_1(K) = \int vK(v)dv$ and $\mu_2(K) = \int v^2 K(v)dv$. For higher d the expression of the equivalent kernel can also be derived, but this would require some more calculations.

The martingale nature of the problem transfers weak convergence into $\mathcal{L}_2$–convergence (see Andersen et al. (1993)), when we assume that the second condition in assumption (iv) holds. Thus we get that

$$\mathrm{E}\left[\widehat{\alpha}_{\mathcal{K},\underline{b}}(x) - \alpha(x)\right]^2 = B_x^2 + (n|\underline{b}|)^{1/2} V_x + o\left(\sum_{j=0}^{d} b_j^4 + (n|\underline{b}|)^{1/2}\right).$$

Applying Fubini's Theorem therefore leads to the following asymptotic expansion of the Mean Integrated Square Errors (MISEs)

$$\begin{aligned} MISE\left(\widehat{\alpha}_{\mathcal{K},\underline{b}}(x)\right) = \;& \textstyle\sum_{j=0}^{d} \kappa_{1j} b_j^2 \int_0^1 \left\{\frac{1}{2}\frac{\partial^2 \alpha(x)}{\partial x_j^2}\right\}^2 dx \\ & + (n|\underline{b}|)^{-1} \kappa_2 \textstyle\int_0^1 \frac{\alpha(x)}{\varphi(x)} dx + o\left(\sum_{j=0}^{d} b_j^4 + (n|\underline{b}|)^{1/2}\right), \end{aligned} \tag{A.1}$$

involving the symmetric multiplicative kernel $\mathcal{K}$, where κ_2 and κ_{1j}, $j = 0, \ldots, d$ are defined above. Similarly, we can write the expression for $MISE\left(\widehat{\alpha}_{\mathcal{K}_A,\underline{b}}(x)\right)$ for the case of asymmetric $\mathcal{K}_A$ analogously to (A.1) by replacing κ_2 and κ_{1j} with $\bar{\kappa}_2$ and $\bar{\kappa}_{1j}$ for $j = 0, \ldots, d$.

The above MISE calculations are required to develop the do-validation method of Mammen et al. (2011). The key is to find the correct rescaling constant C that allows us to transform the MISE optimal bandwidth for the local linear estimator using asymmetric kernels $\mathcal{K}_A$ to be a MISE optimal bandwidth for the estimator using the kernel with symmetric components $\mathcal{K}$. In the paper by Mammen et al. (2011) this constant C was derived by first obtaining closed-form expressions for the MISE-optimal bandwidths for each local linear estimator. However, such closed-form expressions are not available for dimensions $d > 1$ unless we assume that the vector of bandwidths is determined by a single scalar $b > 0$. Instead we start by assuming a simpler situation to get the MISE-optimal bandwidths and calculate C and later we prove that in the general case of $\underline{b} = (b_0, \ldots, b_d)$ this C remains exactly the same. For $\underline{b} = b \times (1, \ldots, 1)$ with $b > 0$, the MISE-optimal b is given by:

$$b_{MISE} = \left\{ \frac{\kappa_2}{\kappa_1^2} \frac{\int \alpha(x)\{\varphi(x)\}^{-1} dx}{\sum_{j=0}^{d} \int (\frac{\partial^2 \alpha(x)}{\partial x_j^2})^2 dx} \right\}^{1/(d+5)} n^{-1/(d+5)}, \tag{A.2}$$

where we assume a symmetric multiplicative kernel with equal components $K_j = K$ for all $j = 0, \ldots, d$. Similarly, the MISE-optimal bandwidth for the estimator $\widehat{\alpha}_{\mathcal{K}_A}$, involving the asymmetric kernel $\mathcal{K}_A$, is given by:

$$b^A_{MISE} = \left\{ \frac{\bar{\kappa}_2}{\bar{\kappa}_1^2} \frac{\int \alpha(x)\{\varphi(x)\}^{-1} dx}{\sum_{j=0}^{d} \int (\frac{\partial^2 \alpha(x)}{\partial x_j^2})^2 dx} \right\}^{1/(d+5)} n^{-1/(d+5)}. \tag{A.3}$$

Therefore the rescaling constant C which is defined as the ratio between (A.2) and (A.3) becomes:

$$C = \left(\frac{\kappa_2/\kappa_1^2}{\bar{\kappa}_2/\bar{\kappa}_1^2} \right)^{1/(d+5)} . \tag{A.4}$$

Now we assume the general case of $\underline{b} = (b_0, \ldots, b_d)$ but with the restriction that $K_j = K$ for all $j = 0, \ldots, d$. Let $\underline{b}^A_{MISE}$ denote the MISE-optimal bandwidth for the estimator $\widehat{\alpha}_{\mathcal{K}_A}$. Through some simple calculations we can prove that the rescaled $C\underline{b}^A_{MISE}$ with $C = \left(\frac{\kappa_2/\kappa_1^2}{\bar{\kappa}_2/\bar{\kappa}_1^2} \right)^{1/(d+5)}$ is optimal in the MISE sense for the estimator $\widehat{\alpha}_{\mathcal{K}}$. We can therefore conclude that for any consistent estimator $\widehat{\underline{b}}^A$ of the optimal $\underline{b}^A_{MISE}$, the rescaled version $C\widehat{\underline{b}}^A$ will be a consistent estimator of the optimal $\underline{b}_{MISE}$ for the preferred hazard estimator $\widehat{\alpha}_{\mathcal{K}}$.

Appendix B. Simulation and approximation strategies

The purpose of this section is to derive the discrete version of the hazard estimator and the corresponding scores defined for bandwidth selection purposes in the the paper. These expressions are necessary to deal with data applications such as mortality studies, where the data are given as occurrences and exposures for discrete points. These occurrences and exposures are just discretised versions of the counting processes N_i and Y_i, defined in Subsection 2. Deriving an expression of the estimator from discretised variables is a simple exercise that we will carry out in Subsection Appendix B.1.

Apart from being required to deal with data applications, the discrete versions are useful for computational purposes in simulation studies. Instead of simulating the continuous scenario formulated in the model directly, it is much faster and more convenient to simulate the data on a lower level of aggregation and then to implement the methods using the discrete versions. Specific algorithms to simulate this kind of data are provided in Subsections Appendix B.2 and Appendix B.3. Note that this strategy in practice is justified as a numerical approximation of the integrals involved in the estimators. Thus, the level of aggregation should be low enough to provide stable approximations. In the simulation experiments described in Section 4 we used a grid of 100 two dimensional points, which was enough to guarantee the stability of the results.

Appendix B.1. Discrete approximations of the hazard estimators and the bandwidth selection strategies

In the continuous scenario we assume that we observe n individuals and have observations of type $\{(N_1, Z_1, Y_1), \ldots, (N_n, Z_n, Y_n)\}$ which are independent and identically distributed. In a practical scenario such detailed data may be provided in an aggregated way. To simplify the derivations presented here we assume that the marker Z is unidimensional ($d = 1$). Let us assume a practical situation where we only have observations at $M < n$ discrete time points, $t_1, \ldots, t_M$, and $M' < n$ discrete values of the covariate, $Z_1, \ldots, Z_{M'}$. Then for each $r' = 1, \ldots, M'$ we have a sample path $\{(N_{r',1}, Y_{r',1}, \ldots, (N_{r',n_{r'}}, Y_{r',n_{r'}})\}$ of the counting process with intensity process $\lambda_k(t, Z_{r'}) = \alpha(t, Z_{r'})Y_{r',l}(t)$, $l = 1, \ldots, n_{r'}$ and $\sum_{r'=1}^{M'} n_{r'} = n$. In this situation the local linear estimator at any estimation point $x = (t, z)^t$ given in (3) can be approximated as follows:

$$\begin{aligned}\widehat{\alpha}_{K,\underline{b}}(x) =& \frac{\sum_{r'=1}^{M'}\sum_{l=1}^{n_{r'}}\sum_{r=1}^{M}\int_{t_{r-1}}^{t_r}[1-(x-x_{r,r'})^t D(x)^{-1}c_1(x)]K_{b_0}(t-s)K_{b_1}(z-Z_{r'})dN_{r',l}(s)}{\sum_{r'=1}^{M'}\sum_{l=1}^{n_{r'}}\sum_{r=1}^{M}\int_{t_{r-1}}^{t_r}[1-(x-x_{r,r'})^t D(x)^{-1}c_1(x)]K_{b_0}(t-s)K_{b_1}(z-Z_{r'})Y_{r',l}(s)ds} \\ \approx& \frac{\sum_{r'=1}^{M'}\sum_{r=1}^{M}[1-(x-x_{r,r'})^t D(x)^{-1}c_1(x)]K_{b_0}(t-t_r^*)K_{b_1}(z-Z_{r'})\sum_{l=1}^{n_{r'}}\int_{t_{r-1}}^{t_r}dN_{r',l}(s)}{\sum_{r'=1}^{M'}\sum_{r=1}^{M}[1-(x-x_{r,r'})^t D(x)^{-1}c_1(x)]K_{b_0}(t-t_r^*)K_{b_1}(z-Z_{r'})\sum_{l=1}^{n_{r'}}\int_{t_{r-1}}^{t_r}Y_{r',l}(s)ds}\end{aligned}$$

with $x_{r,r'} = (t_r, z_{r'})$ and $t_r^* = (t_{r-1} + t_r)/2$, for $r = 1, \ldots, M$ and $r' = 1, \ldots, M'$. Here the bandwidth is $\underline{b} = (b_0, b_1)^t$ and we assume the same kernel K for both the marker and time. From the expression above we can identify the occurrences and the exposures as follows:

$$O_{r,r'} = \sum_{j=1}^{n_{r'}}\int_{t_{r-1}}^{t_r} dN_{r',j}(s),$$

and

$$E_{r,r'} = \sum_{j=1}^{n_{r'}}\int_{t_{r-1}}^{t_r} Y_{r',j}(s)ds,$$

for $r = 1, \ldots, M$ and $r' = 1, \ldots, M'$. Note that $O_{r,r'}$ represents the observed occurrences of the counting processes $\{N_{r',1}, \ldots, N_{r',n_{r'}}\}$, and $E_{r,r'}$ represents the observed exposures from the counting processes $\{Y_{r',1}, \ldots, Y_{r',n_{r'}}\}$ in the interval $[t_{r-1}, t_r)$ (for $r = 1, \ldots, M$ and $r' = 1, \ldots, M'$). Thus, from a common discrete dataset consisting of $\{(x_{r,r'} = (t_r, z_{r'}), E_{r,r'}, O_{r,r'}); r = 1, \ldots, M, r' = 1, \ldots, M'\}$, the local linear hazard estimator can be calculated from the following discrete approximation:

$$\widetilde{\alpha}_{K,\underline{b}}(x) = \frac{\sum_{r'=1}^{M'}\sum_{r=1}^{M}\{1-(x-x_{r,r'})^t\widetilde{D}(x)^{-1}\widetilde{c}_1(x)\}K_{b_0}(t-t_r^*)K_{b_1}(z-Z_{r'})O_{r,r'}}{\sum_{r'=1}^{M'}\sum_{r=1}^{M}\{1-(x-x_{r,r'})^t\widetilde{D}(x)^{-1}\widetilde{c}_1(x)\}K_{b_0}(t-t_r^*)K_{b_1}(z-Z_{r'})E_{r,r'}}, \qquad \text{(B.1)}$$

where $\widetilde{c}_1(x) = (\widetilde{c}_{1,0}(x), \widetilde{c}_{1,1}(x))^t$ and $\widetilde{D}(x) = (\widetilde{d}_{j,k}(x))_{j,k}$ $(j, k = 0, 1)$ are approximations of the moments $c_1(x)$ and $D(x)$ in (3), which are given by:

$$\begin{aligned}
\widetilde{c}_{1,0}(x) &= \sum_{r=1}^{M}\sum_{r'=1}^{M'} K_{b_0}(t-t_r^*)K_{b_1}(z-z_{r'})(t-t_r^*)E_{r,r'} \\
\widetilde{c}_{1,1}(x) &= \sum_{r=1}^{M}\sum_{r'=1}^{M'} K_{b_0}(t-t_r^*)K_{b_1}(z-z_{r'})(z-z_{r'})E_{r,r'} \\
\widetilde{d}_{0,0}(x) &= \sum_{r=1}^{M}\sum_{r'=1}^{M'} K_{b_0}(t-t_r^*)K_{b_1}(z-z_{r'})(t-t_r^*)^2E_{r,r'} \\
\widetilde{d}_{0,1}(x) &= \sum_{r=1}^{M}\sum_{r'=1}^{M'} K_{b_0}(t-t_r^*)K_{b_1}(z-z_{r'})(t-t_r^*)(z-z_{r'})E_{r,r'} \\
\widetilde{d}_{1,1}(x) &= \sum_{r=1}^{M}\sum_{r'=1}^{M'} K_{b_0}(t-t_r^*)K_{b_1}(z-z_{r'})(z-z_{r'})^2E_{r,r'}.
\end{aligned}$$

Using similar arguments one can define discrete versions of the performance measure in (4) and the cross-validation score as follows:

$$Q_0(\underline{b}) \approx n^{-1}\sum_{r=1}^{M}\sum_{r'=1}^{M'}\left[\widetilde{\alpha}_{K,\underline{b}}(x_{r,r'}) - \alpha(x_{r,r'})\right]^2 E_{r,r'}, \tag{B.2}$$

and

$$\widehat{Q}_0(\underline{b}) \approx \sum_{r=1}^{M}\sum_{r'=1}^{M'}\left(\widetilde{\alpha}_{K,\underline{b}}(x_{r,r'})\right)^2 E_{r,r'} - 2\sum_{r=1}^{M}\sum_{r'=1}^{M'}\widetilde{\alpha}_{K,\underline{b}}^{[r,r']}(x_{r,r'})O_{r,r'}, \tag{B.3}$$

where $\widetilde{\alpha}_{K,\underline{b}}^{[r,r']}$ is the (discrete) hazard estimator arising when the dataset is changed by setting $O_{r,r'} = O_{r,r'} - 1$.

Appendix B.2. Simulating complete samples

Here we describe an algorithm to simulate complete samples from the models defined in (7). As we indicated above, the data are obtained in an aggregated way so that each simulated sample will be a set such that $\{(x_{r,r'} = (t_r, z_{r'}), E_{r,r'}, O_{r,r'}); r = 1, \dots, M, r' = 1, \dots, M'\}$. The models are defined with Z being a unidimensional marker with range $(0, 1)$ and time is defined in $[0, 1]$ for models 1 and 2 and in $(0, 5]$ for the two last models. Let us now consider the time interval $[0, \tau]$. The algorithm to simulate a complete (not filtered) sample can be described as follows:

Algorithm.

Step 1. Define the grid points $\{t_r : r = 1, \dots, M\}$ and $\{z_{r'} : r' = 1, \dots, M'\}$ as regular sequences between $[0, \tau]$ and $[0, 1]$, respectively. Let δ_M be the step length of the time sequence i.e. $\delta_M = t_2 - t_1$.

Step 2. Draw n i.i.d. markers $Z_1, \dots, Z_n$ from a Uniform in $[0, 1]$.

Step 3. For each marker value in the grid $z_{r'}$, $r' = 1, \ldots, M'$, the occurrences and the exposures, $\{O_{r,r'}, E_{r,r'}; r = 1, \ldots, m\}$, are derived as follows:

(i) Calculate the number of individuals at risk at the initial time t_1 by counting the number of simulated Z_i falling into $[z_{r'-1}, z_{r'})$. Let $n_{r'}$ denote the total number of observations for each r' (clearly $\sum_{r'=1}^{M'} n_{r'} = n$). Using the notation from subsection Appendix B.1 the discretised risk processes satisfy: $Y_{r'}^{(n_{r'})}(t_1) := \sum_{l=1}^{n_{r'}} Y_{r',l}(t_1) = n_{r'}$.

(ii) Generate the number of failures or occurrences at the initial time t_1, which is denoted by $O_{1,r'}$. This is done by simulating from a Binomial with size $Y_{r'}^{(n_{r'})}(t_1)$ and probability $\alpha(t_1, z_{r'})\delta_M$.

(iii) Calculate the size of the risk set, $Y_{r'}^{(n_{r'})}(t_r)$, and generate the occurrences, $O_{r,r'}$, at the rest of the time points in the grid t_r ($r = 2, \ldots, M$) by the following recursive expressions:

$$\begin{aligned} Y_{r'}^{(n_{r'})}(t_r) &= Y_{r'}^{(n_{r'})}(t_{r-1}) - O_{r-1,r'}, \\ O_{r,r'} &\hookrightarrow \mathrm{B}\left(Y_{r'}^{(n_{r'})}(t_r), \alpha(t_r, z_{r'})\delta_M\right). \end{aligned}$$

(iv) Finally, the exposures $E_{r,r'}$ is calculated as $E_{r,r'} = Y_{r'}^{(n_{r'})}(t_r)\delta_M$.

Step 4. Repeat Step 3 for each $r' = 1, \ldots, M'$ to get the sample $\{(x_{r,r'} = (t_r, z_{r'}), E_{r,r'}, O_{r,r'}); r = 1, \ldots, M, r' = 1, \ldots, M'\}$.

Appendix B.3. Simulating left truncation and right censoring

We will now introduce truncation and censoring following a similar method as Buch-Kromann and Nielsen (2012). For the models 1 to 4 in (7), the number of censored observations was generated from a Binomial distribution with success probability $p_{cens} = 0.01$. This corresponds to an overall proportion of right censoring of about 20% with a grid size of $M = 100$ time points. The proportion of left truncated observations in the sample was $p_{trun} = 0.25$. The algorithm to simulate one sample with these levels of censoring and truncation follows a similar structure to the algorithm in subsection Appendix B.2. The difference is just in Step 3, which now runs as follows:

Step 3. For each marker value in the grid $z_{r'}$, $r' = 1, \ldots, M'$ the occurrences and exposures, $\{O_{r,r'}, E_{r,r'}; r = 1, \ldots, m\}$, are derived as follows:

(i) Calculate the number of simulated Z_i which fall into $[z_{r'-1}, z_{r'})$ and let this number be denoted by $n_{r'}$ ($\sum_{r'=1}^{M'} n_{r'} = n$). Calculate the number of individuals at risk at the initial time t_1 as: $Y_{r'}^{(n_{r'})}(t_1) = n_{r'}(1 - p_{trun}) := \widetilde{n}_{r'}$.

(ii) Generate $\widetilde{n}_{r'}$ random values $\{T_{1,r'}, \ldots, T_{\widetilde{n}_{r'},r'}\}$ from a Uniform in $[0, \tau/2]$. Calculate the number of truncated observations that enter the sample at each time t_r ($r = 1, \ldots, M$) by counting the number of simulated $T_{\cdot,r'}$ ($r = 1, \ldots, M$) which fall into $[t_{r-1}, t_r)$. Let $T_{r,r'}$ denote the resulting counts and update the number of individuals at risk at the initial time t_1 by calculating $Y_{r'}^{(n_{r'})}(t_1) = Y_{r'}^{(n_{r'})}(t_1) + T_{1,r'}$.

(iii) Generate the number of failures or occurrences at the initial time t_1, which is denoted by $O_{1,r'}$. This is done by simulating from a binomial with size $Y_{r'}^{(n_{r'})}(t_1)$ and probability $\alpha(t_1, z_{r'})\delta_M$.

(iv) Generate the censoring level at time t_1 from a binomial with size $Y_{r'}^{(n_{r'})}(t_1) - O_{1,r'}$ and probability p_{cens}. Let the the generated value be denoted by $C_{1,r'}$

(v) Calculate the size of the risk set ($Y_{r'}^{(n_{r'})}(t_r)$) and generate the occurrences ($O_{r,r'}$) and the censoring levels ($C_{r,r'}$) at the rest of time points in the grid t_r ($r = 2, \ldots, M$) by the following recursive expressions:

$$\begin{aligned} Y_{r'}^{(n_{r'})}(t_r) &= Y_{r'}^{(n_{r'})}(t_{r-1}) - O_{r-1,r'} + C_{r-1,r'} - T_{r-1,r'}, \\ O_{r,r'} &\hookrightarrow \mathrm{B}\left(Y_{r'}^{(n_{r'})}(t_r), \alpha(t_r, z_{r'})\delta_M\right), \\ C_{r,r'} &\hookrightarrow \mathrm{B}\left(Y_{r'}^{(n_{r'})}(t_{r-1}) - O_{r-1,r'}, p_{cens}\right). \end{aligned}$$

(vi) Finally, the exposure $E_{r,r'}$ is calculated as $E_{r,r'} = Y_{r'}^{(n_{r'})}(t_r)\delta_M$.