Journal of Machine Learning Research 20 (2019) 1-29 Submitted 11/17; Published 2/19

Multiplicative local linear hazard estimation and best one-sided cross-validation

María Luz Gámiz MGAMIZ@UGR.ES
Department of Statistics and Operations Research, University of Granada
Campus Fuentenueva s/n, 18071 Granada, Spain

María Dolores Martínez-Miranda MMIRANDA@UGR.ES
Department of Statistics and Operations Research, University of Granada
Campus Fuentenueva s/n, 18071 Granada, Spain

Jens Perch Nielsen JENS.NIELSEN.1@CITY.AC.UK
Cass Business School, City, University of London
106 Bunhill Row, London EC1Y8TZ, U.K.

Editor: Sara van de Geer

Abstract

This paper develops detailed mathematical statistical theory of a new class of cross-validation techniques of local linear kernel hazards and their multiplicative bias corrections. The new class of cross-validation combines principles of local information and recent advances in indirect cross-validation. A few applications of cross-validating multiplicative kernel hazard estimation do exist in the literature. However, detailed mathematical statistical theory and small sample performance are introduced via this paper and further upgraded to our new class of best one-sided cross-validation. Best one-sided cross-validation turns out to have excellent performance in its practical illustrations, in its small sample performance and in its mathematical statistical theoretical performance.

Keywords: Aalen's multiplicative model, multiplicative bias correction, bandwidth, indirect cross-validation

1. Introduction

There is a growing interest in validation techniques. While validation was always a crucial element of mathematical statistics, the use of validation techniques are growing rapidly at the moment under labels such as big data, machine learning or artificial intelligence. Many of these developments seem less patient with laborious mathematical statistical model formulation and estimation theory than what has been the trademark of the field of mathematical statistics. Instead inspiration seems to be taken from neighbouring fields such as engineering, computer science, public health or actuarial science, where specific knowledge is present on the problem at hand, allowing the development of clever and perhaps computationally challenging algorithms often replacing more labour intensive procedures of the past. These algorithms are often defined in such a way that they can change and learn over time via some optimization criteria and an efficient validation procedure. One example of such work relevant to the work of this paper is Muñoz and van der Laan (2012) where an impressive algorithm is developed to solve a complicated survival problem. The introduced methodology is inspired by machine learning: the validation procedure is called a Super Learner. However, while the Super Learner is optimal in some sense, see van der Laan et al. (2007), then it is not optimal in the more detailed mathematical statistical sense that we consider in this paper. And this is not only because Muñoz and van der Laan (2012) consider piecewise constant hazard models that are less efficient than kernel smoothers. It is also because the validation theory provided by Muñoz and van der Laan (2012) does not provide the mathematical detail promoted in this paper, and therefore crucial insight of noisy second order components is not included in the theory. The approach of Muñoz and van der Laan (2012) is just one among many machine learning inspired survival analyses approaches. This paper considers one dimension only. Multidimensional cross-validation and one-dimensional cross-validation are closely related and mathematical definitions are similar. However, even in the one-dimensional case we face challenging theoretical, as well as practical issues, with cross-validation being too noisy and unstable and to such an extend that we cannot any longer recommend cross-validation in one dimension without some amendment for the noise involved. Our intention is that multidimensional big data type of problems, with further issues with data sparsity and noisy cross-validation, should benefit in the future from the insight on cross-validation analyses as provided in this paper. The mathematical point of view of this paper was perhaps initiated via the early contribution of Hall and Marron (1987) that provided a decision theoretical framework to distinguish between plug-in estimators, aiming at minimizing a mean integrated squared error, and cross-validation aiming at minimizing the infeasible stochastic integrated squared error. They concluded that plug-in did better from an asymptotic perspective even when the aim was the explicit aim of cross-validation: to get as close as possible to the infeasible minimization of the integrated squared error. One could view this as the foundation of a new decision theoretical framework to understand the quality of kernel bandwidth selection; a tractable place to start when understanding the complicated world of model selection. Hall and Johnstone (1992) pointed out that for any bandwidth selector there are two sources of noise for kernel density estimation, one that one can never get rid of and another one that seems to differ for different methods. The second source of noise can theoretically go as low as to zero such that one was left with the first noise component as a lower bound on noise. The plug-in type of methods had considerable lower second-component noise than cross-validation and plug-in was very popular in practice in the nineties, with Sheather and Jones (1991) being the perhaps most popular single method. However, plug-in methods depend on complicated underlying mathematical detail that does not easily generalize to new problems in the same straightforward way as cross-validation does. This is perhaps the single most important reason why cross-validation has regained its importance and is used for a wide variety of complicated problems in mathematical statistics, big data, machine learning and artificial intelligence. Hart and Yi (1998) introduced the concept of indirect cross-validation, formulated in nonparametric kernel regression, which simply meant that cross-validation was performed on an alternative kernel and the bandwidth was scaled back to the original kernel used for estimation. Hart and Yi (1998) suggested to use one-sided kernels as the alternative kernels because of their good practical performance and simple rescaling. In density estimation Savchuk et al. (2010) suggested a combination of a normal-bandwidth kernel and an oversmoothed kernel as alternative kernel to achieve the same mathematical statistical asymptotic performance as the plug-in estimator without the need of a pilot. However, there was one catch with the elegant approach of Savchuk et al. (2010). Their approach needed to estimate some tuning parameters to decide the relative weight of the oversmoothed kernel that was contributing to the asymptotic noise via some term of lower order. So, even though Savchuk et al. (2010) in principle did pilot free estimation then there was still some tuning going on and some extra terms of just slightly lower order. And that was perhaps exactly the problem of the original plug-in methods as in Sheather and Jones (1991): that something with lower order noise had to be estimated, the pilot, and terms of slightly lower order had to be ignored in the asymptotic results. In this paper we define three dogmas for a modern kernel smoothing estimator:

1. It should be a direct estimation based on principles without complicated mathematical adjustments.

2. Extra terms of slightly lower order are not allowed in the expansions.

3. Further smoothing than those necessary for the original estimator is not allowed to be assumed while analysing the quality of the bandwidth selector.

The original cross-validation estimator and the approach of Hart and Yi (1998) lives up to all three dogma rules while the plug-in type estimators of e.g. Sheather and Jones (1991) and Savchuk et al. (2010) violate all three. We believe this to be the reason why Mammen et al. (2011,2014) concluded that their double one-sided kernel density bandwidth selector, directly inspired by Hart and Yi (1998), worked better in practice than the estimators of Sheather and Jones (1991) and Savchuk et al. (2010). The fundamental principles of this paper are therefore the three dogmas above and the decision theoretical framework of Hall and Marron (1987), and this has let us to explore double one-sided cross-validation and one-sided cross-validation even further because of their apparent practical superiority on the market of current kernel bandwidth selectors. A detailed investigation of both sides of local one-sided bandwidth selection showed us a perhaps surprising fact. While the left-sided and the right-sided cross-validation procedures have the same mathematical statistical behaviour, they do perform very differently in practice. Often one of the two sides breaks down completely. Double one-sided cross-validation works better than one-sided cross-validation in a wide variety of kernel smoothing problems, see for example Mammen et al. (2011,2014), Gámiz et al. (2013a,b, 2016). A closer investigation going through local features of individual simulation samples reveals that behind a good double one-sided cross-validation result often hides an average of a good one-sided estimator and a somehow random result from the other side. The suggestion of this paper is to improve the stability of one-sided cross-validation via a local information principle inspecting at every single local point whether to use the right side or the left side for cross-validation. This approach is stable in its practical performance, it obeys the three above dogmas and it provides the exact same asymptotic performance as the less stable one-sided and double-sided competitors mentioned above. We call the new approach best one-sided cross-validation. This paper furthermore introduces the mathematical statistical approach of Hall and Marron (1987) to multiplicative bias corrected local linear kernel hazard estimators and it introduces asymptotic theory and practical implementation of best one-sided-cross-validation for these multiplicative bias corrected hazard estimators. Multiplicative bias correction is known to improve the practical implementation of kernel hazard estimation, see Nielsen

(1998) and Nielsen and Tanggaard (2001). This parallels insights from the more researched world of kernel density estimation, see for example Jones et al. (1995) and Jones and Signorini (1997). The latter went through a series of small sample studies of kernel density estimation procedures to conclude that multiplicative bias correction seemed to be the best. The contribution of this paper is therefore also to update mathematical statistical theory and practice to the perhaps best practically performing kernel hazard estimator we have: the multiplicative bias corrected local linear kernel hazard estimator.

The rest of the paper is organized as follows. In Section 2 we describe the link between our proposal and methods in machine learning. In Section 3 we formulate the model we assume in the paper and present two hazard estimators, namely the local linear estimator and its multiplicative bias correction. Bandwidth selection for these estimators through cross-validation and the double one-sided cross-validation method of Gámiz et al. (2016) is described in Section 4, and our new best one-sided cross-validation method is suggested. The asymptotic properties of all presented validated bandwidths are analysed in Section 5. Assumptions are deferred to Appendix A and proofs are provided in the Supplementary Material. A further investigation of the theoretical properties of bandwidth selectors is described in Appendix B. Two case studies show the applicability of our proposals, which are described in Section 6. In Section 7 we describe simulation experiments to evaluate the finite sample properties of our proposal. The main findings from the simulations are discussed in Section 8 along with further insights about the asymptotic properties of bandwidth selectors. Final conclusions are drawn in Section 9. All numerical calculations have been performed with `R` and the methods proposed in this paper have been implemented in the `DOvalidation` package (Gámiz et al., 2017).

2. Training and learning versus cross-validation and adjusted cross-validation

To motivate our research beyond a wider crowd than experts in nonparametric hazard estimation, our point of view is formulated below via standard vocabulary from machine learning and artificial intelligence. Let us assume we observe n individuals over some time that could potentially be filtered via truncation and censoring, and let A be a training set and B be a learning set such that the two sets united equals the set $\{1, \ldots, n\}$. Let for the purpose of a discussion the number of elements of A be 80% of n and the number of elements in B be 20% of n. Then a standard approach to validation, see again Muñoz and van der Laan (2012), would be to estimate the hazard on the training set and evaluate it via the learning set. Under some standard independence assumptions this will lead to a decrease in efficiency of estimation itself corresponding to ignoring 20% of the data set and it will decrease the efficiency on the validation approach, compared to cross-validation and the theoretical approach considered in this paper, corresponding to ignoring 80% of the data set. One could of course consider all possible combinations of training and learning sets and average all these validations into one single validation principle or learning principle. This would correspond to a computationally inefficient cross-validation. In conclusion: even if all possible combinations of trainers and learners are calculated, we end up with standard cross-validation with the well known problems of data sparsity and noise. With the help of the theory originally developed by Hall and Marron (1987) in the kernel density context, we will in this paper, in the kernel hazard context, consider more efficient use of data when estimators are validated or when trainers are learning. It turns out that this is indeed possible via relatively straightforward adjustments of standard cross-validation.

3. The counting process model and kernel hazard estimators

In this section we formulate events via counting processes. Counting processes are well designed when event data are filtered for example via truncation or censoring. An individual zero-one valued exposure process simply keeps tracks on whether an individual is under risk or not at any particular point in time. We assume that individuals are independent and that data filtering is non-informative. Formally, 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 an increasing, right continuous, complete filtration $\mathcal{F}_t$, $t \in [0, T]$, i.e., it obeys the usual conditions (Andersen et al., 1993, p. 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 $\alpha(\cdot)$. Here Y_i is a predictable process taking values 0 or 1, indicating (by the value 1) when the ith individual is at risk and under observation. We assume that $(N_1, Y_1), \ldots, (N_n, Y_n)$ are independent and identically distributed for the n individuals. With these definitions λ_i is predictable and the processes $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, \ldots, n$, with $\Lambda_i(t) = \int_0^t \lambda_i(s)\ ds$, are square integrable local martingales.

As an example we illustrate how the above stochastic processes look like in the case of independent and non-informative left truncation and right censoring, where n tuples (L_i, Z_i, δ_i), $i = 1, \ldots, n$, are observed. Here L_i is the time the ith individual enters the study; Z_i is the time ith individual leaves the study either because an event has happened or because of right censoring; and δ_i is binary and equal to one if an event, for example death or an onset of a disease, is the reason for the ith individual to leave the study, and the value is zero when the reason for the ith individual to leave the study was uninformative right censoring. In this case, 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 convenient and general stochastic process formulation only.

The local linear kernel hazard estimator in our general stochastic process formulation was introduced by Nielsen and Tanggaard (2001) and it is defined as

$$\widehat{\alpha}_{b,K}^{\mathrm{LL}}(t) = \sum_{i=1}^{n} \int_0^T \bar{K}_{t,b}(t-s) dN_i(s), \tag{1}$$

with the stochastic local linear kernel

$$\bar{K}_{t,b}(t-s) = \frac{a_{2,K}(t) - a_{1,K}(t)(t-s)}{a_{0,K}(t)a_{2,K}(t) - \{a_{1,K}(t)\}^2} K_b\left(t-s\right), \tag{2}$$

where $K_b(u) = b^{-1}K(u/b)$, $a_{j,K}(t) = \int_0^T K_b\left(t-s\right)(t-s)^j Y(s)ds$, for $j = 0, 1, 2$, and $Y(t) = \sum_{i=1}^n Y_i(t)$ is the aggregated risk process. Here K is a kernel function with support $[-1, 1]$ and $b > 0$ is the bandwidth parameter.

The local linear kernel $\bar{K}_{t,b}$ satisfies the properties: $\int_0^T \bar{K}_{t,b}(t-s)Y(s)ds = 1$, $\int_0^T \bar{K}_{t,b}(t-s)(t-s)Y(s)ds = 0$ and $\int_0^T \bar{K}_{t,b}(t-s)(t-s)^2 Y(s)ds > 0$. Thus, $\bar{K}_{t,b}$ can be interpreted as a second order kernel with respect to the stochastic measure μ, where $d\mu(s) = Y(s)ds$. Defining the aggregated failure process, $N(t) = \sum_{i=1}^{n} N_i(t)$, we can write $\widehat{\alpha}_{b,K}^{\text{LL}}(t) = \int_0^T \bar{K}_{t,b}(t-s)dN(s)$.

The multiplicative bias corrected estimator constructed from the local linear hazard estimator is defined as

$$\widehat{\alpha}_{b,K}^{\text{MBC}}(t) = \sum_{i=1}^{n} \int \bar{K}_{t,b}^{\text{MBC}}(t-s)\widehat{\alpha}_{b,K}^{\text{LL}}(t)\{\widehat{\alpha}_{b,K}^{\text{LL}}(s)\}^{-1}dN_i(s), \tag{3}$$

where the multiplicative kernel is

$$\bar{K}_{t,b}^{\text{MBC}}(t-s) = \frac{a_{2,K}^{\text{MBC}}(t) - a_{1,K}^{\text{MBC}}(t)(t-s)}{a_{0,K}^{\text{MBC}}(t)a_{2,K}^{\text{MBC}}(t) - \{a_{1,K}^{\text{MBC}}(t)\}^2} \left\{\widehat{\alpha}_{b,K}^{\text{LL}}(s)\right\}^2 K_b\left(t-s\right), \tag{4}$$

with $a_{j,K}^{\text{MBC}}(t) = \int_0^T K_b\left(t-s\right)(t-s)^j \left\{\widehat{\alpha}_{b,K}^{\text{LL}}(s)\right\}^2 Y(s)ds$, for $j = 0, 1, 2$.

4. Cross-validation and best one-sided cross-validation of our two estimators

The two kernel hazards estimators considered in this paper depend on a bandwidth parameter that determines the smoothness degree of the resulting estimates. Choosing the bandwidth parameter is a crucial problem that starts by defining what the optimal bandwidth would be, so it can be estimated from data.

Let $\widehat{\alpha}_{b,K}$ denote a kernel hazard estimator with bandwidth b and kernel K, which can be any of the two defined in (1) or (3). Ideally we would like a bandwidth parameter b that minimizes the integrated squared error (ISE) defined as

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

where $w(\cdot)$ is some weight function. However, the minimizer of the integrated squared error, $\widehat{b}_{\text{ISE},K}$, depends on the unknown hazard function and it is infeasible in practice. In this paper we consider $\widehat{b}_{\text{ISE},K}$ as the optimal bandwidth and in this section we present estimates based on the cross-validation method. We refer the reader to Gámiz et al. (2016) for the history of cross-validation in kernel hazard estimation based on counting processes.

First notice that minimizing $\Delta_K(b)$ is equivalent to minimizing

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

and only the second term depends on the unknown hazard. The cross-validation approach estimates this second term from the data replacing $\alpha(s)ds$ by its empirical counterpart $dN_i(s)$. The cross-validated bandwidth, denoted by $\widehat{b}_{\text{CV},K}$, is therefore the minimizer of

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

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

A practical and theoretical improvement of cross-validation was given in Gámiz et al. (2016) that developed double one-sided cross-validation (DO-validation), as a simple average of two indirect cross-validated bandwidths. 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. Double one-sided cross-validation averages the two indirect cross-validation bandwidths based on one-sided kernels: the left-sided $K_{\mathrm{L}}(u) = 2K(u)I(u < 0)$, or the right-sided $K_{\mathrm{R}}(u) = 2K(u)I(u > 0)$. More specifically, two one-sided cross-validation criteria, $\widehat{Q}_{K_{\mathrm{L}}}(b)$ and $\widehat{Q}_{K_{\mathrm{R}}}(b)$, are defined as in (5) but replacing K with K_{L} and K_{R}, respectively. Denoting by $\widehat{b}_{\mathrm{CV},K_{\mathrm{L}}}$ and $\widehat{b}_{\mathrm{CV},K_{\mathrm{R}}}$ their minimizers, the double one-sided cross-validation bandwidth estimate is the (conveniently) weighted average of these

$$\widehat{b}_{\mathrm{DO},K} = \rho \left\{ \frac{\widehat{b}_{\mathrm{CV},K_{\mathrm{L}}} + \widehat{b}_{\mathrm{CV},K_{\mathrm{R}}}}{2} \right\}.$$

For the local linear hazard estimator defined in (1), the factor ρ is

$$\rho^{\mathrm{LL}} = \left\{ \frac{R(K)}{R(\bar{K}^*_{\mathrm{L}})} \frac{\mu_2(\bar{K}^*_{\mathrm{L}})^2}{\mu_2(K)^2} \right\}^{1/5}. \tag{6}$$

Here, for a general kernel L, $\bar{L}^*$ denotes the equivalent local linear kernel defined as

$$\bar{L}^*(u) = \frac{\mu_2(L) - \mu_1(L)u}{\mu_2(L) - \mu_1(L)^2} L(u), \tag{7}$$

where $\mu_l(L) = \int u^l L(u)du$, for $l = 1, 2$, and $R(L) = \int L^2(u)du$. Notice that $\bar{L}^* = L$ if L is symmetric.

For the multiplicative bias corrected estimator, $\widehat{\alpha}^{\mathrm{MBC}}_{b,K}$, defined in (3), the factor ρ becomes

$$\rho^{\mathrm{MBC}} = \left\{ \frac{R(\Gamma_K)}{R(\Gamma_{\bar{K}^*_{\mathrm{L}}})} \frac{\mu_2(\bar{K}^*_{\mathrm{L}})^4}{\mu_2(K)^4} \right\}^{1/9}, \tag{8}$$

where $\Gamma_L(u) = 2L(u) - L(u) * L(u)$ is the kernel obtained by twicing the kernel L. Here $*$ denotes the convolution operator.

The asymptotic theory developed in Gámiz et al. (2016) for the local linear hazard estimator showed that left- and right-sided cross-validation have the same asymptotic properties, but different finite sample performance. There are situations where one of the two one-sided cross-validation methods breaks down so the averaging strategy of double one-sided cross-validation becomes inappropriate. The natural reaction in these cases would be to take the side which is working fine. One common reason for one of the two one-sided cross-validated bandwidths to break down is the lack of occurrences (or exposures) in one of the two directions. Best one-sided cross-validation (BO-validation), introduced in this paper simply uses the one-sided version that, via local information, is predicted to work best at every single point t. There can therefore be both left-sided and right-sided kernels involved in best one-sided cross-validation. Imagine for example that the estimation interval is $(0,1)$, where two boundaries are present, then one would expect to use different sided kernels for a t close to the left boundary 0 and for a t close to the right boundary 1.

For the local linear hazard estimator we define the kernel estimator needed for best one-sided cross-validation as

$$\widehat{\alpha}_{b,K}^{\mathrm{BO,LL}}(t) = \int_0^T \left[\bar{K}_{t,b;\mathrm{L}}(t-s)\xi_b(t) + \bar{K}_{t,b;\mathrm{R}}(t-s)\left\{1-\xi_b(t)\right\}\right] dN(s), \tag{9}$$

where $\bar{K}_{t,b;\mathrm{L}}$ and $\bar{K}_{t,b;\mathrm{R}}$ are respectively the left and right versions of the local linear kernel $\bar{K}_{t,b}$ in (2), and $\xi_b(t)$ is a stochastic function, depending on the estimation time t and the bandwidth b, which takes the value 1 when the "best" side to consider is the indicated by the kernel K_{L}, and the value 0 otherwise. The combination of one-sided kernels that appears in the integrand of expression (9) is a kernel function which we denote as

$$\bar{K}_{b,K}^{\mathrm{BO,LL}}(t-s) = \bar{K}_{t,b;\mathrm{L}}(t-s)\xi_b(t) + \bar{K}_{t,b;\mathrm{R}}(t-s)\left\{1-\xi_b(t)\right\}. \tag{10}$$

Thus we write the estimator as $\widehat{\alpha}_{b,K}^{\mathrm{BO,LL}}(t) = \int_0^T \bar{K}_{b,K}^{\mathrm{BO,LL}}(t-s)dN(s)$.

For each time t, to designate which side is "best", $\xi_b(t)$ can be defined in terms of the occurrence process by

$$\xi_b^{\mathrm{O}}(t) = I\left(\int_{t-b}^{t} dN(s) < \int_t^{t+b} dN(s)\right),$$

or the exposure process by

$$\xi_b^{\mathrm{E}}(t) = I\left(\int_{t-b}^{t} Y(s)ds < \int_t^{t+b} Y(s)ds\right). \tag{11}$$

With any of these ξ_b^{O} or ξ_b^{E}, the best one-sided cross-validation bandwidth estimate is defined as

$$\widehat{b}_{\mathrm{BO},K}^{\mathrm{LL}} = \rho^{\mathrm{LL}} \arg\min_b \widehat{Q}_K^{\mathrm{BO,LL}}(b), \tag{12}$$

where $\widehat{Q}_K^{\mathrm{BO,LL}}$ is the cross-validation score in (5) calculated with the kernel estimator $\widehat{\alpha}_{b,K}^{\mathrm{BO,LL}}(t)$, defined in (9). In a similar way we define the best one-sided cross-validation bandwidth estimate for the multiplicative bias corrected estimator, $\widehat{b}_{\mathrm{BO},K}^{MBC}$, as in (12) but replacing the factor ρ^{LL} with ρ^{MBC}, given in (8), and defining the best one-sided cross-validation score, $\widehat{Q}_K^{\mathrm{BO,MBC}}$, with the hazard estimator

$$\widehat{\alpha}_{b,K}^{\mathrm{BO,MBC}}(t) = \int_0^T \left[\bar{K}_{t,b;\mathrm{L}}^{\mathrm{MBC}}(t-s)\frac{\widehat{\alpha}_{b,K_{\mathrm{L}}}^{\mathrm{LL}}(t)}{\widehat{\alpha}_{b,K_{\mathrm{L}}}^{\mathrm{LL}}(s)}\xi_b(t) + \bar{K}_{t,b;\mathrm{R}}^{\mathrm{MBC}}(t-s)\frac{\widehat{\alpha}_{b,K_{\mathrm{R}}}^{\mathrm{LL}}(t)}{\widehat{\alpha}_{b,K_{\mathrm{R}}}^{\mathrm{LL}}(s)}\left\{1-\xi_b(t)\right\}\right] dN(s). \tag{13}$$

5. Asymptotic theory

In this section we develop theory for the asymptotic behaviour of bandwidth selectors for the local linear hazard estimator in (1), and its multiplicative bias correction in (3). For each estimator we prove the asymptotic normality for bandwidths selectors based on cross-validation, the double one-sided cross-validation of Gámiz et al. (2016) and the new best one-sided cross-validation. Our theoretical results thus extend the results given in Gámiz et al. (2016), by including the new best one-sided cross-validation for local linear hazard estimator and considering its multiplicative bias correction.

Recall that the integrated squared error of a kernel hazard estimator, $\widehat{\alpha}_{b,L}$, with bandwidth b and kernel L, was defined as above as

$$\Delta_L(b) = n^{-1} \int_0^T \left\{ \widehat{\alpha}_{b,L}(t) - \alpha(t) \right\}^2 w(t) Y(t) dt, \tag{14}$$

and its minimizer denoted as $\widehat{b}_{\mathrm{ISE},L}$. Hereafter we will make explicit reference to the considered hazard estimator using superscripts (LL for the local linear and MBC for the multiplicative bias correction). Besides a kernel denoted by K is assumed to be symmetric, while we use the notation L for a general kernel that can be asymmetric, as the one-sided kernels involved in double one-sided cross-validation and best one-sided cross-validation (see assumption A1 in Appendix A).

Let consider first the local linear hazard estimator, $\widehat{\alpha}^{\mathrm{LL}}_{b,L}$, given in (1). Following the same arguments described in Nielsen and Tanggaard (2001), the error $\widehat{\alpha}^{\mathrm{LL}}_{b,L}(t) - \alpha(t)$, can be decomposed as, $\widehat{\alpha}^{\mathrm{LL}}_{b,L}(t) - \alpha(t) = V^{\mathrm{LL}}_{b,L}(t) + B^{\mathrm{LL}}_{b,L}(t)$, where $B^{\mathrm{LL}}_{b,L}$ is a stable part converging in probability to zero,

$$B^{\mathrm{LL}}_{b,L} = \int_0^T \bar{L}_{t,b}(t-s) \left\{ \alpha(s) - \alpha(t) \right\} Y(s) ds; \tag{15}$$

and $V^{\mathrm{LL}}_{b,L}$ is a variable part converging to a Normal distribution,

$$V^{\mathrm{LL}}_{b,L}(t) = \int_0^T \bar{L}_{t,b}(t-s) dM(s). \tag{16}$$

Using the above decomposition we can expand the integrated squared error for the local linear estimator, using standard martingale theory along with the approach of Mammen and Nielsen (2007). In Lemma 4 in the Supplementary Material we show that, under some regularity assumptions, $\Delta^{\mathrm{LL}}_L(b)$ in (14) is asymptotically equivalent to

$$M^{\mathrm{LL}}_L(b) = b^4 \frac{\mu_2(\bar{L}^*)^2}{4} \int \left\{ \alpha''(t) \right\}^2 \gamma(t) w(t) dt + (nb)^{-1} R\left(\bar{L}^*\right) \int \alpha(t) w(t) dt,$$

where $\gamma(t) = n^{-1} E\left[Y(t) \right]$ is the expected exposure function. From this approximation a deterministic optimal bandwidth for the local linear estimator with kernel L is defined as

$$b^{\mathrm{LL}}_{\mathrm{MISE},L} = C^{\mathrm{LL}}_{0,L} n^{-1/5} \qquad \text{with} \qquad C^{\mathrm{LL}}_{0,L} = \left[\frac{R\left(\bar{L}^*\right) \int \alpha(t) w(t) dt}{\mu_2(\bar{L}^*)^2 \int \left\{ \alpha''(t) \right\}^2 \gamma(t) w(t) dt} \right]^{1/5}. \tag{17}$$

Our main result in this section states the asymptotic normality of the three bandwidth estimates for the local linear hazard estimator, cross-validation, $\widehat{b}^{\mathrm{LL}}_{\mathrm{CV},K}$, double one-sided cross-validation, $\widehat{b}^{\mathrm{LL}}_{\mathrm{DO},K}$, and best one-sided cross-validation, $\widehat{b}^{\mathrm{LL}}_{\mathrm{BO},K}$; as well as the infeasible bandwidth $\widehat{b}^{\mathrm{LL}}_{\mathrm{MISE},K}$. Note that the latter is the optimal bandwidth targeted by plug-in bandwidth selection rules. The proof of the theorem is provided in the Supplementary Material.

Theorem 1 *Under assumptions A1–A3 in Appendix A, the bandwidth selectors, $\widehat{b}^{\mathrm{LL}}_{\mathrm{BO},K}$, $\widehat{b}^{\mathrm{LL}}_{\mathrm{DO},K}$,$\widehat{b}^{\mathrm{LL}}_{\mathrm{CV},K}$, and $\widehat{b}^{\mathrm{LL}}_{\mathrm{MISE},K}$, for the local linear estimator with kernel K satisfy*

$$\begin{aligned}
n^{3/10}\left(\widehat{b}^{\mathrm{LL}}_{\mathrm{BO},K}-\widehat{b}^{\mathrm{LL}}_{\mathrm{ISE},K}\right) &\longrightarrow N\left(0, S_2^{\mathrm{LL}}+S_1^{\mathrm{LL}}\Psi^{\mathrm{LL}}_{\mathrm{BO},K}\right)\\
n^{3/10}\left(\widehat{b}^{\mathrm{LL}}_{\mathrm{DO},K}-\widehat{b}^{\mathrm{LL}}_{\mathrm{ISE},K}\right) &\longrightarrow N\left(0, S_2^{\mathrm{LL}}+S_1^{\mathrm{LL}}\Psi^{\mathrm{LL}}_{\mathrm{DO},K}\right)\\
n^{3/10}\left(\widehat{b}^{\mathrm{LL}}_{\mathrm{CV},K}-\widehat{b}^{\mathrm{LL}}_{\mathrm{ISE},K}\right) &\longrightarrow N\left(0, S_2^{\mathrm{LL}}+S_1^{\mathrm{LL}}\Psi^{\mathrm{LL}}_{\mathrm{CV},K}\right)\\
n^{3/10}\left(\widehat{b}^{\mathrm{LL}}_{\mathrm{MISE},K}-\widehat{b}^{\mathrm{LL}}_{\mathrm{ISE},K}\right) &\longrightarrow N\left(0, S_2^{\mathrm{LL}}+S_1^{\mathrm{LL}}\Psi^{\mathrm{LL}}_{\mathrm{MISE},K}\right)
\end{aligned}$$

where

$$S_1^{\mathrm{LL}} = \frac{1}{25}\frac{R(K)^{-7/5}\int \alpha^2(t)w^2(t)\ dt}{\mu_2(K)^{6/5}\left\{\int \alpha''(t)^2\gamma(t)w(t)\ dt\right\}^{3/5}\left\{\int \alpha(t)w(t)\ dt\right\}^{7/5}},$$

$$S_2^{\mathrm{LL}} = \frac{4}{25}\frac{R(K)^{-2/5}\int \alpha''(t)^2\gamma(t)w^2(t)\alpha(t)\ dt}{\mu_2(K)^{6/5}\left\{\int \alpha(t)w(t)\ dt\right\}^{2/5}\left\{\int \alpha''(t)^2\gamma(t)w(t)\ dt\right\}^{8/5}},$$

and

$$\begin{aligned}
\Psi^{\mathrm{LL}}_{\mathrm{BO},K} &= \Psi^{\mathrm{LL}}_{\mathrm{DO},K} = \int\left\{\frac{R(K)}{R(\bar{L}^*)}\left(H_L-G_L\right)\left(\rho^{\mathrm{LL}}u\right)-H_K(u)\right\}^2 du,\\
\Psi^{\mathrm{LL}}_{\mathrm{CV},K} &= \int\left\{G_K(u)\right\}^2\ du,\\
\Psi^{\mathrm{LL}}_{\mathrm{MISE},K} &= \int\left\{H_K(u)\right\}^2\ du,
\end{aligned}$$

defining the functions $G_L(\cdot)$ and $H_L(\cdot)$ as

$$\begin{aligned}
G_L(w) &= I\left(w\neq 0\right)2\bar{L}_1^*(w),\\
H_L(w) &= I\left(w\neq 0\right)\int \bar{L}^*(u)\left\{\bar{L}_1^*(u+w)+\bar{L}_1^*(u-w)\right\}du,
\end{aligned}$$

where $\bar{L}_1^*(u)=-\bar{L}^*(u)-u\bar{L}^{*\prime}(u)$, with $L=K$ and $L=K_{\mathrm{L}}$.

Remark 2 *Gámiz et al. (2016) pointed out that all bandwidth estimates have similar asymptotics with the only difference of the factor $\Psi^{\mathrm{LL}}_{\cdot,K}$. These authors considered three*

Table 1: Comparison of asymptotic variances among bandwidth selectors. Factors $\Psi^{\bullet}_{\bullet,K}$ defined in Theorems 1 and 3, are shown for the local linear hazard estimator and its multiplicative bias correction, with three common symmetric kernels K: Epanechnikov, quartic and sextic.

	Local linear estimator			Multiplicative bias correction		
Method	Epanechnikov	Quartic	Sextic	Epanechnikov	Quartic	Sextic
BO-validation	1.09	0.95	1.18	4.41	2.44	2.05
DO-validation	1.09	0.95	1.18	4.41	2.44	2.05
Cross-validation	3.60	2.86	3.49	9.87	6.10	6.50
Plug-in	0.36	0.46	0.59	0.84	0.95	1.31

common choices of the kernel K (Epanechnikov, quartic and sextic kernels) and calculated the numerical value of this factor (multiplied by 2 for convenience in this former paper). It allows the comparison of the asymptotic performance of bandwidth selectors. These numerical values are reported in the first rows of Table 1, where now we have added the new best one-sided cross-validation. Notice that the actual values of S_1^{LL} and S_2^{LL} depend on the unknown hazard function α and its derivatives, but also on the exposure function γ, which are specific to the inference problem. In Appendix B.1 we have calculated the ratio of these two terms for one particular hazard model. The idea is to find out whether the discrepancy in the values of $\Psi^{\mathrm{LL}}_{\cdot,K}$ among bandwidth selectors shown in the table could be misleading in terms of the actual variances. For the considered model the term S_1^{LL} is almost a half of S_2^{LL}, both when exposure time goes zero and infinity. For the multiplicative bias corrected hazard estimator we will discuss in Remark 4 that the corresponding S_1 term can completely dominate the term S_2 in certain situations.

Consider now the multiplicative bias correction of the local linear hazard defined in (3), $\widehat{\alpha}^{\mathrm{MBC}}_{b,L}$, with bandwidth b and kernel L. As for the local linear estimator above, we define the corresponding integrated squared error for its multiplicative bias correction as in (14) and denote it as $\Delta_L^{\mathrm{MBC}}(b)$. We denote its minimizer as $\widehat{b}^{\mathrm{MBC}}_{\mathrm{ISE},L}$.

We consider the decomposition $\widehat{\alpha}^{\mathrm{MBC}}_{b,L}(t)-\alpha(t) = B^{\mathrm{MBC}}_{b,L}(t)+V^{\mathrm{MBC}}_{b,L}(t)$, where $B^{\mathrm{MBC}}_{b,L}(t)$ is a stable term converging in probability to zero, and $V^{\mathrm{MBC}}_{b,L}(t)$ is a variable term converging to a Normal distribution. These two terms are defined as follows:

$$V^{\mathrm{MBC}}_{b,L}(t) = \int f^{\mathrm{MBC}}_{t,b}(s)dM(s)$$

where

$$\begin{aligned} f^{\mathrm{MBC}}_{t,b}(s) \quad &= \quad \bar{L}^{\mathrm{MBC}}_{t,b}(t-s)\frac{\widehat{\alpha}^{\mathrm{LL}}_{b,L}(t)}{\widehat{\alpha}^{\mathrm{LL}}_{b,L}(s)} + \bar{L}_{t,b}(t-s) - \\ &\int_0^T \bar{L}^{\mathrm{MBC}}_{t,b}(t-u)\frac{\widehat{\alpha}^{\mathrm{LL}}_{b,L}(t)}{\widehat{\alpha}^{\mathrm{LL}}_{b,L}(u)}\bar{L}_{u,b}(u-s)Y(u)du \end{aligned}$$

with $\bar{L}^{\mathrm{MBC}}_{t,b}(t-s)$ defined as in (4) for the kernel L, and

$$B_{b,L}^{\mathrm{MBC}}(t) = B_{b,L}^{\mathrm{LL}}(t) + \int \bar{L}_{t,b}^{\mathrm{MBC}}(t-s)\widehat{\alpha}_{b,L}^{\mathrm{LL}}(t)\left\{\widehat{\alpha}_{b,L}^{\mathrm{LL}}(s)\right\}^{-1} B_{b,L}^{\mathrm{LL}}(s)Y(s)ds$$
$$= \int \bar{L}_{t,b}^{\mathrm{MBC}}(t-s)\widehat{\alpha}_{b,L}^{\mathrm{LL}}(t)\left\{\beta_{b,L}(t) - \beta_{b,L}(s)\right\} Y(s)ds,$$

with $\beta_{b,L}(s) = \{\widehat{\alpha}_{b,L}^{\mathrm{LL}}(s)\}^{-1} B_{b,L}^{\mathrm{LL}}(s)$, where $B_{b,L}^{\mathrm{LL}}$ and $V_{b,L}^{\mathrm{LL}}$ are the stable and variable terms for the local linear estimator given in (15) and (16), respectively.

Using the above decomposition and derivations similar to the local linear case we can expand the integrated squared error of the multiplicative bias corrected estimator. In Lemma 7 in the Supplementary Material we show that, under some regularity assumptions, $\Delta_L^{\mathrm{MBC}}(b)$ is asymptotically equivalent to

$$M_L^{\mathrm{MBC}}(b) = b^8 \frac{\mu_2(\bar{L}^*)^4}{16} \int h(t)^2 \gamma(t) w(t) dt + (nb)^{-1} R\left(\Gamma_{\bar{L}^*}\right) \int \alpha(t) w(t) dt,$$

with $h(t) = \alpha(t)\left\{\alpha''(t)/\alpha(t)\right\}''$. From this approximation a deterministic optimal bandwidth for the multiplicative bias corrected estimator with kernel L is defined as

$$b_{\mathrm{MISE,L}}^{\mathrm{MBC}} = C_{0,L}^{\mathrm{MBC}} n^{-1/9}; \qquad C_{0,L}^{\mathrm{MBC}} = \left\{ \frac{R\left(\Gamma_{\bar{L}^*}\right) \int \alpha(t) w(t) dt}{\frac{\mu_2(\bar{L}^*)^4}{2} \int h(t)^2 \gamma(t) w(t) dt} \right\}^{1/9}, \tag{18}$$

where $\Gamma_{\bar{L}^*}(u) = 2\bar{L}^*(u) - \bar{L}^*(u) * \bar{L}^*(u)$ is the kernel obtained by twicing the equivalent kernel, $\bar{L}^*$, given in (7).

The following theorem states the asymptotic normality of the three bandwidth estimates, as well as the minimizer of the mean integrated squared error, for the multiplicative bias corrected hazard estimator with kernel K. The proof is provided in the Supplementary Material.

Theorem 3 *Under assumptions A1, A2' and A3', the bandwidth selectors* $\widehat{b}_{\mathrm{BO},K}^{\mathrm{MBC}}$, $\widehat{b}_{\mathrm{DO},K}^{\mathrm{MBC}}$, $\widehat{b}_{\mathrm{CV},K}^{\mathrm{MBC}}$, *and* $\widehat{b}_{\mathrm{MISE},K}^{\mathrm{MBC}}$ *satisfy*

$$\begin{aligned}
n^{3/18}\left(\widehat{b}_{\mathrm{BO},K}^{\mathrm{MBC}} - \widehat{b}_{\mathrm{ISE},K}^{\mathrm{MBC}}\right) &\longrightarrow N\left(0, S_2^{\mathrm{MBC}} + S_1^{\mathrm{MBC}} \Psi_{\mathrm{BO},K}^{\mathrm{MBC}}\right)\\
n^{3/18}\left(\widehat{b}_{\mathrm{DO},K}^{\mathrm{MBC}} - \widehat{b}_{\mathrm{ISE},K}^{\mathrm{MBC}}\right) &\longrightarrow N\left(0, S_2^{\mathrm{MBC}} + S_1^{\mathrm{MBC}} \Psi_{\mathrm{DO},K}^{\mathrm{MBC}}\right)\\
n^{3/18}\left(\widehat{b}_{\mathrm{CV},K}^{\mathrm{MBC}} - \widehat{b}_{\mathrm{ISE},K}^{\mathrm{MBC}}\right) &\longrightarrow N\left(0, S_2^{\mathrm{MBC}} + S_1^{\mathrm{MBC}} \Psi_{\mathrm{CV},K}^{\mathrm{MBC}}\right)\\
n^{3/18}\left(\widehat{b}_{\mathrm{MISE},K}^{\mathrm{MBC}} - \widehat{b}_{\mathrm{ISE},K}^{\mathrm{MBC}}\right) &\longrightarrow N\left(0, S_2^{\mathrm{MBC}} + S_1^{\mathrm{MBC}} \Psi_{\mathrm{MISE},K}^{\mathrm{MBC}}\right)
\end{aligned}$$

where

$$S_1^{\mathrm{MBC}} = \frac{2^{1/3}}{9^2} \frac{R\left(\Gamma_K\right)^{-5/6} \int \alpha(t)^2 w(t)^2 \, dt}{\mu_2(K)^{4/3} \left\{\int h(t)^2 \gamma(t) w(t) \, dt\right\}^{1/3} \left\{\int \alpha(t) w(t) \, dt\right\}^{5/3}},$$

$$S_2^{\mathrm{MBC}} = \frac{2^{10/3}}{9^2} \frac{R\left(\Gamma_K\right)^{-2/3} \int h(t)^2 \gamma(t) w(t)^2 \alpha(t) \, dt}{\mu_2(K)^{4/3} \left\{\int \alpha(t) w(t) \, dt\right\}^{2/3} \left\{\int h(t)^2 \gamma(t) w(t) \, dt\right\}^{4/3}},$$

with

$$\begin{aligned}
\Psi^{\mathrm{MBC}}_{\mathrm{BO},K} &= \Psi^{\mathrm{MBC}}_{\mathrm{DO},K} = \int \left\{ \frac{R\left(\Gamma_K\right)}{R\left(\Gamma_{\bar{L}^*}\right)} \left(H_{\Gamma_{\bar{L}^*}} - G_{\Gamma_{\bar{L}^*}}\right)\left(\rho^{\mathrm{MBC}} u\right) - H_{\Gamma_K}(u) \right\}^2 du, \\
\Psi^{\mathrm{MBC}}_{\mathrm{CV},K} &= \int \left\{G_{\Gamma_K}(u)\right\}^2 \, du, \\
\Psi^{\mathrm{MBC}}_{\mathrm{MISE},K} &= \int \left\{H_{\Gamma_K}(u)\right\}^2 \, du.
\end{aligned}$$

where the functions $G_L(\cdot)$ and $H_L(\cdot)$ are defined as in Theorem 1, with $L = \Gamma_K$ and $L = \Gamma_{\bar{L}^*}$, defined above.

Remark 4 *The result above shows that all bandwidth selectors have similar asymptotics with the only difference of the factor $\Psi^{\mathrm{MBC}}_{\cdot,K}$. A similar conclusion was derived for the local linear estimator. The three last columns of Table 1 show the value of this factor for three common choices of K. As for the local linear estimator we have calculated the ratio $S_1^{\mathrm{MBC}}/S_2^{\mathrm{MBC}}$ for a particular hazard model, the results are shown in Appendix B.2. In this case the term S_1^{MBC}, relative to the term S_2^{MBC}, is negligible when exposure time goes to zero, but dominates completely (infinite times bigger) when exposure time goes to infinity.*

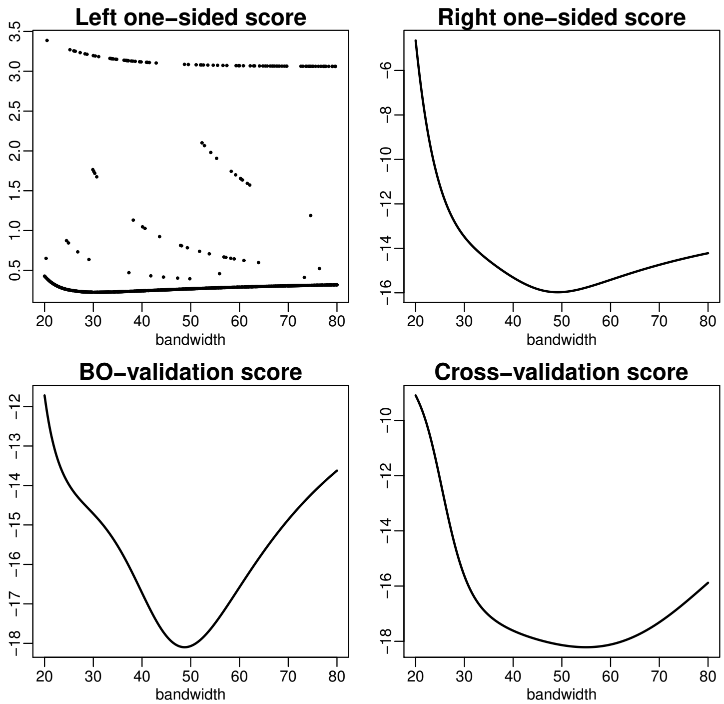
6. Applications

6.1. Old-age mortality

Our first application is on fitting hazard mortality curves for old-age population. We consider mortality data of women in Iceland in the calendar year 2006, with ages from 40 to 110. The same data were considered by Gámiz et al. (2016) and are available in the `DOvalidation` R-package (Gámiz et al., 2017). This package provides also functions implementing the hazard estimators and the bandwidth selection methods described above. The data were obtained from the Human Mortality Database and consist of aggregated yearly occurrences and exposures. Gámiz et al. (2016) showed that estimating the hazard from these data is challenge at the oldest ages. The lack of exposure at the right end and the few observed deaths induce a marked boundary effect precisely in the area of interest, the old ages. For these data we have calculated the two hazard estimators described in this paper, local linear and its multiplicative bias correction, using three bandwidth selectors: cross-validation, double one-sided cross-validation and the new best one-sided cross-validation. The cross-validation scores involved in these methods have been defined using a weighting function such that $w(s)Y(s) \equiv 1$, so all points in the time interval where the hazard function is estimated are evaluated with the same weight. This is different from Gámiz et al. (2016) where the weighting function was chosen so only areas where the exposure is significant contribute to the criteria. Notice that this makes an important difference in this data set where the end of the time interval comprises almost no exposure.

Before looking at the resulting hazard estimates we shall look at the cross-validation scores to be minimized for each bandwidth selection method. Figure 1 shows the cross-validated scores for each method considering the multiplicative bias corrected estimator. The local linear case looks quite similar and can be found in the Supplementary Material. From these plots we can see that the left one-sided score is not well behaved for both

Figure 1: Mortality data: bandwidth selection scores with multiplicative bias corrected hazard estimator.

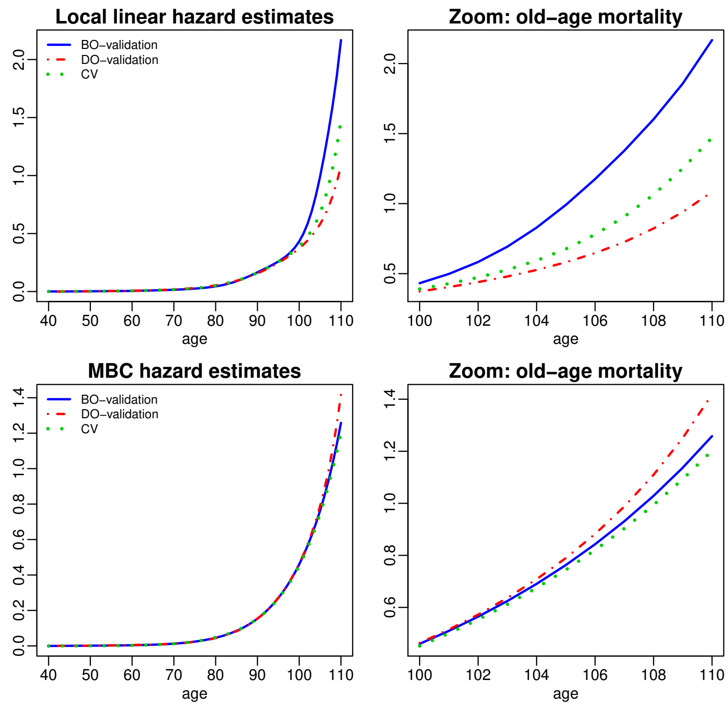


hazard estimators. Therefore the average DO-validated bandwidth becomes unreliable, even though the obtained values seem to be sensible ($\widehat{b}_{\mathrm{DO}} = 27.3$ for the local linear estimator and $\widehat{b}_{\mathrm{DO}} = 40$ for its multiplicative bias correction). On the other hand the best one-sided cross-validation method shows a clear minimum in both cases and, as expected, it moves close to the one-sided cross-validated bandwidth that is working fine (the right side in this case). Best one-sided cross-validation in this case has been calculated using the exposure process, that is, for each time t we use the function $\xi_b^E(t)$ given in (11). However the results are quite similar using the occurrence process instead. Figure 2 shows the resulting hazard estimates from each method and type of hazard estimate. Note from these plots that the multiplicative bias corrected hazard is more robust to the bandwidth choice than the local linear. Also the new best one-sided cross-validation method seems to provide a reasonable estimate for old-age mortality in both cases.

6.2. Prediction of outstanding liabilities in non-life insurance

We consider now a non-standard forecasting problem that arises in non-life insurance. The goal is to forecast the number of future claims from contracts underwritten in the past, which have not yet been reported. Typically actuaries are responsible of getting these forecasts, which represent perhaps the most important number in the accounts of the company (see

Figure 2: Comparison of hazard estimates from female mortality data in Iceland.



Martínez-Miranda et al. (2013) for a detailed background of this problem). Here we analyse a data set of reported and outstanding claims from a motor business in UK. The same data set was previously considered by Martínez-Miranda et al. (2013) and consists of $n = 1558$ large claims reported between January 1990 and March 2012. From a statistical perspective the data could be described as a sample $\{(X_1, Z_1), \ldots, (X_n, Z_n)\}$, where X_i denotes the underwriting date of the ith claim, and Z_i the corresponding reporting delay, this is, the time between the underwriting date and the reporting date of the claim. The sample is right truncated since it can be observed only those claims for which the underwriting time plus the reporting delay is not greater than the calendar time of data collection. Hence data exist on a triangle with $X_i + Z_i \leq 31$ March 2012, and $X_i + Z_i$ represents the calendar time. The aim is to forecast the mass of the unobserved, future triangle, where $X_i + Z_i > 31$ March 2012, which corresponds to the number of claims underwritten in the past which have not been reported yet. The problem is formulated assuming that the maximum reporting delay is 267 months, in the actuarial literature this assumption is described as the triangle is fully run off. Another challenge of the data set for this problem is that the data are only available in an aggregated way. This is a common feature of this kind of data in the reserving departments of the insurance companies. This means that the available observations are counts living in a triangle of dimension 267×267. Specifically for our data set the triangle has entries $\mathcal{N}_{x,z} = \sum_{i=1}^{n} I(X_i = x,\ Z_i = z)$, with $x, z \in \{1, \ldots, 267\}$, describing the number of claims underwritten in the xth month and reported in the zth month.

Martínez-Miranda et al. (2013) showed that a multiplicative structured density model, $f(x,z) = f_1(x)f_2(z)$, can be used to forecast the claims where the components f_1 and f_2 are the underwriting time density and the reporting time density, respectively. The assumption of a multiplicative density means that the reporting delay does not depend on the underwriting date. Using the counting process formulation considered in this paper, Hiabu et al. (2016) solved the forecasting problem estimating the two density components using a time-reversal approach. Data are transformed to the time reversed scale so the right-truncation problem is replaced by the more tractable left-truncation. Using the same time-reversal approach, we now use the hazard estimation methods presented in the previous sections to estimate the backward hazard functions corresponding to the two components, underwriting (α_1) and reporting delay (α_2). From these hazard estimates the density component estimates can be derived multiplying by respective estimators of the survival functions.

From the above description we solve the forecasting problem considering both local linear hazard estimator and its multiplicative bias correction. For each hazard component, the bandwidth parameters for these estimators have been estimated using cross-validation, double one-sided cross-validation and best one-sided cross-validation. In the three cases we use weighting functions for the involved cross-validation scores that are appropriate for the forecasting problem. Specifically, following the discussion in Hiabu et al. (2016), to estimate α_1 we consider weights $w_1(t) = \widehat{S}_1^2(t)\left\{1 - \widehat{S}_2(t)\right\}^2/Y_1(t)$, where $\widehat{S}_1$ and $\widehat{S}_2$ are estimators of the survival functions of each component (underwriting time and the reporting time delay) on the reversed time scale; and $Y_1(t)$ is the risk process for the first component. In a similar way we define the weights to estimate α_2. As in the mortality study best one-sided cross-validation has been calculated using the exposure process.

Figure 3 shows the forecasts of the number of claims reported in the future calendar months. Table 2 shows these forecasts aggregated in years. The forecasts are given for each hazard estimator and bandwidth estimate. We have also included the forecasts derived from the chain ladder method, which involves histogram type estimators of the underwriting and reporting density components. The chain ladder method is the classical approach used in the insurance companies (see Martínez-Miranda et al. (2013) for more details about this approach). The plot of the forecasts shows that the classical insurance method chain ladder is overestimating the liabilities, while the kernel hazard methods provide lower forecasts. Previous empirical analyses with these data described in Martínez-Miranda et al. (2013) agree with this result and recommend multiplicative bias corrected local linear estimators for this kind of data. Looking at the results from the kernel estimators we can see that double one-sided cross-validation and best one-sided cross-validation provide similar forecasts when the local linear estimator is considered, but the results are quite different for the multiplicative bias corrected estimator. The predicted total number of claims using double one-sided cross-validated bandwidth is about 299, compared to 313 using best one-sided cross-validation. Our concern is that double one-sided cross-validation might not be behaving properly in this situation. A close inspection to the cross-validation scores that are minimized to derive these bandwidth estimates reveals what is happening. Figures 4 and 5 show these cross-validation scores when the multiplicative bias corrected estimator is considered for both underwriting and reporting delay components. From these plots we can see that the right one-sided score completely breaks down for the underwriting time compo-

Table 2: Forecasts of the number of claims to be reported in the future calendar years.

Year	CLM	LL-CV	LL-DO	LL-BO	MBC-CV	MBC-DO	MBC-BO
2012	99.95	76.85	77.98	77.95	80.55	81.75	81.76
2013	97.23	75.06	75.52	76.86	81.18	75.82	81.68
2014	74.32	58.75	59.05	60.04	62.23	58.89	62.88
2015	49.18	38.88	39.06	39.44	40.31	38.81	41.20
2016	24.52	19.42	19.50	19.66	20.01	19.34	20.44
2017	11.61	9.35	9.39	9.44	9.60	9.45	9.76
2018	6.21	5.07	5.06	5.09	5.15	4.99	5.27
2019	3.24	2.54	2.53	2.52	2.52	2.53	2.61
2020	1.36	1.25	1.23	1.23	1.23	1.18	1.22
2021	0.99	1.03	1.02	1.02	0.99	0.95	0.95
2022	1.11	0.85	0.84	0.85	0.87	0.83	0.88
2023	1.06	0.71	0.71	0.73	0.81	0.80	0.85
2024	1.20	0.81	0.81	0.84	0.93	0.90	0.94
2025	1.14	0.91	0.92	0.95	0.97	0.93	0.94
> 2025	1.94	1.59	1.59	1.61	1.51	1.48	1.55
Total	375.07	293.07	295.20	298.23	308.86	298.69	312.92

nent, exhibiting several local minima. For the reporting delay component the score function continues decreasing as the value of the bandwidth increases, so it reaches the minimum at the upper limit of the search interval of bandwidths. The left one-sided score behaves more reasonably for the underwriting component but again breaks down for the reporting delay component. This means that one shouldn't trust the double one-sided cross-validation bandwidth derived from these two one-sided criteria, even though the derived estimates in this case turned to be reasonable values, $\widehat{b}_{\mathrm{DO}} = 55.8$ for the underwriting time, and $\widehat{b}_{\mathrm{DO}} = 31.6$ for the delay. On the contrary, the new best one-sided cross-validation method provides bandwidth estimates of $\widehat{b}_{\mathrm{BO}} = 43.4$ for the underwriting time and $\widehat{b}_{\mathrm{BO}} = 11.8$ for the delay, exhibiting well-behaved minimization scores as shown in Figure 4. Regarding to the cross-validation method it exhibits a rather flat score in the underwriting component leading to the large bandwidth estimate of $\widehat{b}_{\mathrm{CV}} = 63.5$, and a value of $\widehat{b}_{\mathrm{CV}} = 11.6$ for the delay that is close to the best one-sided cross-validated bandwidth. The impact of the cross-validated bandwidths on the forecasts is moderated, about 309 predicted claims compared to the 313 from best one-sided cross-validation. We have performed the same inspection with the local linear estimators. These plots can be seen in the Supplementary Material. The picture is again quite similar showing a poor performance of double one-sided cross-validation, however the impact on the forecasts in this case is not substantial. The total number predicted from cross-validation is about 293, compared to 295 from double one-sided cross-validation and 298 for best one-sided cross-validation.

7. Finite sample performance

In this section we investigate the finite sample performance of the new best one-sided cross-validation method for the local linear hazard estimator and its multiplicative bias correction. We have considered the same five hazard models described in Gámiz et al. (2016) (see also

Figure 3: Number of outstanding claims forecast using the local linear estimator and its multiplicative bias correction.

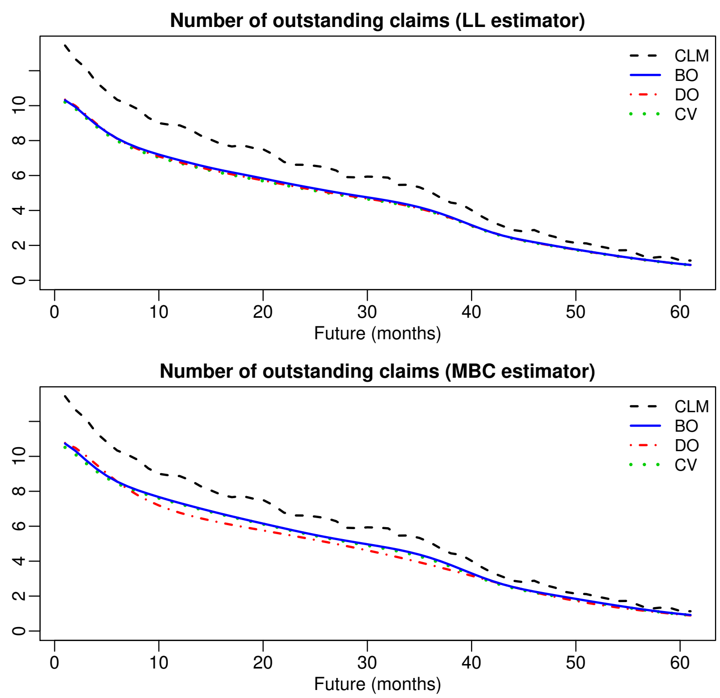


Figure 4: Underwriting component: bandwidth selection scores with multiplicative bias corrected hazard estimator.

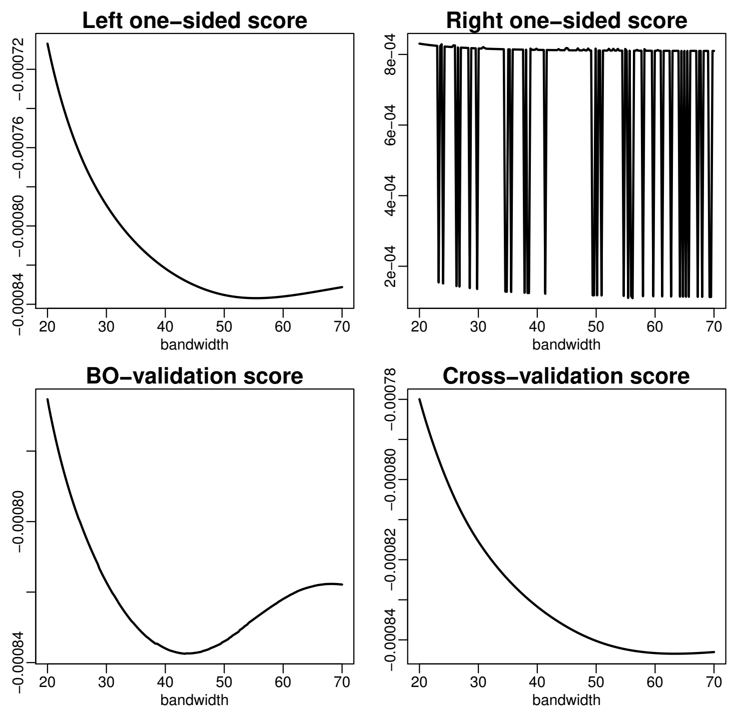
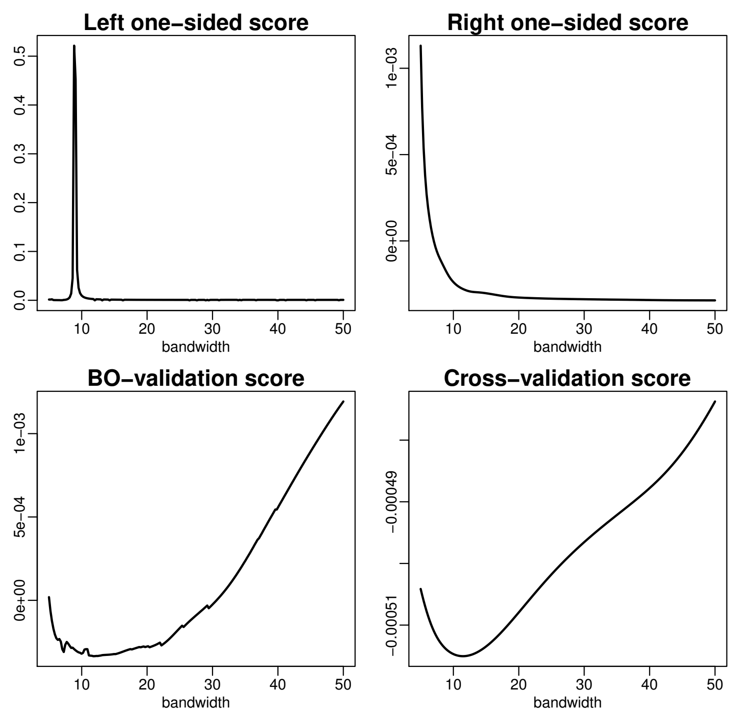


Figure 5: Reporting delay component: bandwidth selection scores with multiplicative bias corrected estimator.



Supplementary Material). The first four models consist of mixtures of Beta densities. Model 5 shows an exponential increase common in hazard mortality rates as those described in the first case study of Section 6. From each model we have simulated samples with three different sample sizes and two sampling schemes, right censoring with and without left truncation. For models 1 to 4 we have considered sample sizes $n = 100, 1000, 10000$, and for model 5, $n = 50000, 75000, 100000$. The number of Monte Carlo replications for each case has been always 500. We use the same mechanism to simulate data as in Gámiz et al. (2016). It generates data in aggregated form (number of occurrences and exposure) for an equally-spaced grid of size R defined on the time interval, and always produces right censored samples. For models 1 to 4 the time interval is $(0, 1)$ and we have defined the grid length with $\delta_R = 1/(R + 1)$. For model 5 time lies in the interval $(40, 110)$ and we have defined the grid length with $\delta_R = 70/(R + 1)$. The grid size has been chosen equal to $R = 500$ in both cases. We shall denote the grid points by t_r $(r = 1, \ldots, R)$. In the case of samples without left truncation, for a sample of n individuals, the number of occurrences at time t_r, denoted as O_r, have been generated from the binomial distribution $Bi \{Y_r, \alpha(t_r)\delta_R\}$, for $r = 1, \ldots, R$. 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 does not sum 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 sample are right censored and the censoring rates are around 20–30% for all models. When adding left truncation, independent truncation times are generated from the Uniform distribution.

From the simulated aggregated data we have calculated the local linear hazard estimator and its multiplicative bias correction using the sextic kernel: $K(x) = 3003/2048(1 - x^2)^6 I(-1 < x < 1)$, as in the two data analyses above. For each hazard estimator we have compared the best one-sided cross-validated bandwidth with cross-validation and double one-sided cross-validation. The performance of the bandwidth estimates have been analysed with respect to the (Monte Carlo approximated) mean integrated squared error of the resulting kernel hazard estimator. We shall refer to this performance measure as empirical MISE, denoted as $m_1(\widehat{b})$, for each bandwidth estimate $\widehat{b}$. As benchmarks in our analysis we have considered two infeasible optimal bandwidths: the bandwidth minimizing the integrated squared error criterion, $\widehat{b}_{\mathrm{ISE}}$, and the bandwidth minimizing the empirical MISE. To compute all bandwidth estimates we have considered grids of 100 equally spaced bandwidth values chosen around $\widehat{b}_{\mathrm{ISE}}$, for each model and sample size. All criteria have been defined using a weighting function such that $w(s)Y(s) \equiv 1$, so all points in the time interval where the hazard function is estimated are evaluated with the same weight. As we pointed out in our first case study this is different from Gámiz et al. (2016), and it makes an important difference in models such as Model 5 where the end of the time interval comprises almost no exposure.

Table 3 summarizes the simulation results in the case of samples with right censoring and left truncation. In this table bandwidth estimates are compared according to measure m_1. For convenience we report a relative measure to indicate when best one-sided cross-validation outperforms cross-validation. The relative measure is defined as:

$$Rerr(BO) = \left\{ m_1(\widehat{b}_{\mathrm{CV}}) - m_1(\widehat{b}_{\mathrm{ISE}}) \right\} / \left\{ m_1(\widehat{b}_{\mathrm{BO}}) - m_1(\widehat{b}_{\mathrm{ISE}}) \right\}. \tag{19}$$

With this definition values of $Rerr(BO)$ above 1 indicate that best one-sided cross-validation outperforms cross-validation. An analogous relative measure, $Rerr(DO)$, has been defined for double one-sided cross-validation. Notice that $Rerr(BO)$ greater than $Rerr(DO)$ indicates that best one-sided cross-validation outperforms double one-sided cross-validation. An overall view of the numbers in the table confirms that best one-sided cross-validation for the multiplicative hazard estimator always outperforms cross-validation, exhibiting $Rerr(BO)$ values above 1, and double one-sided cross-validation for all models except for few cases, where double one-sided cross-validation provides slightly lower empirical MISE values. The results for the local linear estimator show that double one-sided cross-validation and best one-sided cross-validation behave quite similarly, both outperforming in general cross-validation. The case of samples without left truncation is shown in Table 4. It brings similar conclusions though in this case best one-sided cross-validation is beaten by double one-sided cross-validation for Model 5. This case deserves a deeper analysis and it is shown in Table 5. In this table we have shown the empirical MISE defined above and denoted by $m_1(\widehat{b})$, for each bandwidth estimate $\widehat{b}$, as well as the average of the bandwidth estimates for all the samples ($avg(\widehat{b})$), and we have included the left and right one-sided cross-validated bandwidths, from which double one-sided cross-validation is derived. From these results we can clearly see that the left one-sided bandwidth completely breaks down, for all sample sizes and both hazard estimators, while the right side behaves well (notice the large values of the empirical MISE for the left one-sided bandwidth in contrast with those values for the right one-sided bandwidth). The average of the left and right one-sided bandwidths (which double one-sided cross-validation performs) seems to be hiding the problem of the left side, and sometimes it even provides quite reasonable values. Notice that the double one-sided bandwidths are on average closer to the best ISE-optimal bandwidths than the best one-sided cross-validation for the multiplicative bias corrected estimator. However this happens because the double one-sided bandwidth is the average of a small left one-sided bandwidth and a large right one-sided bandwidth. On the other hand best one-sided cross-validation is behaving as the best of the two sides, as we would expect. A similar picture can be seen when analysing the behaviour of double one-sided cross-validation for Model 4 in the case of truncated samples (the full simulation results are provided in the Supplementary Material).

In summary, the simulation results indicate that best-one sided cross-validation and double one-sided cross-validation do better than one-sided cross-validation (that sometimes breaks down) and standard cross-validation. However, it is not always which one is the better of best one-sided cross-validation or double one-sided cross-validation. We suggest to try out both best one-sided and double one-sided cross-validation in any empirical study.

8. Discussion

The overall conclusion of our finite sample section is that double one-sided cross-validation and best one-sided cross-validation are better than standard cross-validation, however, there is no clear winner between double one-sided cross-validation and best one-sided cross-validation. In practice we would suggest to consider both. It is also concluded that the multiplicative bias corrected local linear hazard estimator is more often than not better than the simpler local linear hazard estimator. There is a tendency that double one-sided cross validation is better than best one-sided cross-validation for the simple local linear case,

Table 3: Simulation results for datasets with right censoring and left truncation. Hazard estimators and bandwidth selectors are compared by the relative measure $Rerr(\cdot)$ defined in (19).

Model	n	LL-DO	LL-BO	MBC-DO	MBC-BO
1	100	1.55	1.25	1.47	1.79
	1000	2.32	2.00	0.97	2.88
	10000	1.90	1.71	1.82	3.30
2	100	2.28	2.04	0.46	2.47
	1000	2.42	1.99	0.15	3.66
	10000	2.18	1.84	0.34	3.81
3	100	1.86	1.74	1.47	1.27
	1000	0.96	0.99	0.82	1.19
	10000	2.20	2.07	2.12	3.50
4	100	0.08	1.12	2.13	0.92
	1000	2.51	1.91	2.30	1.08
	10000	2.17	1.83	3.76	2.62
5	50000	1.62	1.70	1.77	2.09
	75000	2.04	2.18	1.41	2.31
	10^5	1.68	1.73	1.07	1.90

Table 4: Simulation results for datasets without left truncation. Hazard estimators and bandwidth selectors are compared by the relative measure $Rerr(\cdot)$ defined in (19).

Model	n	LL-DO	LL-BO	MBC-DO	MBC-BO
1	100	2.58	2.05	0.89	2.51
	1000	2.62	2.27	1.24	4.60
	10000	2.75	2.47	1.62	8.57
2	100	2.55	1.81	0.22	2.92
	1000	2.70	2.29	0.10	3.51
	10000	2.63	2.40	0.26	4.71
3	100	1.50	1.40	0.99	0.70
	1000	2.72	2.33	0.74	3.40
	10000	1.81	2.10	0.65	3.52
4	100	2.03	1.89	2.19	1.13
	1000	2.09	2.03	1.28	0.90
	10000	1.24	1.28	1.03	1.65
5	50000	0.80	6.45	5.33	1.60
	75000	0.63	5.47	4.63	1.96
	10^5	0.56	4.32	4.16	2.28

Table 5: Performance of double one-sided cross-validation in simulations. The empirical MISE (multiplied by 10^6), $m_1(\widehat{b})$, and the average bandwidth estimates, $avg(\widehat{b})$, are shown for samples generated from Model 5 without left truncation.

	n		ISE	MISE	CV	OSCV-l	OSCV-r	DO	BO
LL	50000	$m_1(\widehat{b})$	1.14	2.16	13.31	13424.45	3.04	16.29	3.03
		$avg(\widehat{b})$	9.71	22.48	18.09	2.82	10.99	6.91	10.98
	75000	$m_1(\widehat{b})$	0.82	1.32	8.82	4835.00	1.92	10.28	1.92
		$avg(\widehat{b})$	9.37	20.91	16.18	3.57	9.77	6.67	9.76
	10^5	$m_1(\widehat{b})$	0.53	0.86	4.44	1894.00	1.43	7.55	1.43
		$avg(\widehat{b})$	9.04	20.12	14.34	3.98	8.98	6.48	8.96
MBC	50000	$m_1(\widehat{b})$	0.33	0.72	11.84	203897.80	7.01	2.49	7.54
		$avg(\widehat{b})$	18.93	21.70	45.62	17.69	41.43	29.56	40.78
	75000	$m_1(\widehat{b})$	0.23	0.43	5.74	92670.27	3.01	1.42	3.04
		$avg(\widehat{b})$	18.45	20.91	38.88	18.02	35.20	26.61	32.31
	10^5	$m_1(\widehat{b})$	0.15	0.27	3.51	16302.04	1.65	0.96	1.63
		$avg(\widehat{b})$	17.87	20.12	32.85	19.01	30.79	24.90	27.20

while the performance is vice versa when using the multiplicative kernel hazard estimation. The exception from this rule seems to be our finite sample study in Table 5 inspired by our real-life mortality data.

In Appendix B we have studied how big the second noise component (the one that is improved by our new bandwidth selectors) is compared to the first noise component in the bandwidth selection (the one that is same across all the bandwidth selectors). Using our real-life mortality data as inspiration, we investigate a simple version of a Gomperz-Makeham shaped hazard and conclude that for the local linear hazard the second variance component is almost a half of the first component, both when exposure time goes zero and infinity. This picture is very different for the multiplicative hazard estimator, where the second variance component is negligible when exposure time goes to zero, but dominates completely (infinite times bigger) when exposure time goes to infinity. One can in other words get all kinds of relationships between the second variance component and first variance component. This calls for more research, but it seems already clear that there will be many situations, where we cannot rely on simple cross-validation (not to mention less efficient training-and-learning algorithms), because cross-validation has a very high second variance component and that second variance component will occasionally dominate completely the smoothing selection problem. We believe our findings to be relevant for machine learning, big data and artificial intelligence in general, where over-reliance of simple training-and-learning or cross-validation might lead to false-discoveries because of the noise involved. Methods such as double one-sided cross-validation, best one-sided cross-validation, or other improved smoothing procedures dampening the second noise component, therefore seem to be a very important element of future research in machine learning and related fields.

9. Conclusion

We have proposed a new bandwidth selection method for local linear hazard estimation and its multiplicative bias correction. Our proposal is called best one-sided cross-validation and consists of an improvement of the double one-sided cross-validation of Gámiz et al. (2016). Best one-sided cross-validation solves the lack of stability of double one-sided cross-validation in practice via a local information principle.

Our empirical studies show that best one-sided cross-validation provides a good strategy for bandwidth selection for both local linear and multiplicative bias corrected hazard estimators. Best one-sided cross-validation inherits the good properties of one-sided cross-validation while avoiding the stability problems that double one-sided cross-validation sometimes faces. The current algorithm is only about optimisation of statistical inference. However, it could be also interesting to consider computational performance, see for example Kapotufe and Verma (2017).

Detailed mathematical theory at the level of Hall and Marron (1987) and Gámiz et al. (2016) is included. This type of theory is completely novel for the multiplicative bias corrected hazard estimators. Theory on best one-sided cross-validation introduced in this paper is of course also new for the local linear hazard estimator.

Acknowledgments

The authors are grateful for constructive comments from an anonymous Reviewer and the Associate Editor. This work has been partially supported by the Spanish Ministry of Economy and Competitiveness, through grant number MTM2016-76969P, which include support from the European Regional Development Fund (ERDF). The authors thank Centro de Servicios de Informática y Redes de Comunicaciones (CSIRC), University of Granada, for providing the computing time.

Appendix A. Assumptions for asymptotic theory

Assumption A1. The kernels K and L 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 d such that $|\phi(x) - \phi(y)| \leq c|x-y|^d$ for $x, y < 0$ or $x, y > 0$ with ϕ equal to K' or L'. The left- and right-sided derivatives differ at most on a finite set. The kernel K is symmetric.

Assumption A2. For the expected exposure function $\gamma(t) = n^{-1}E\{Y(t)\}$ it holds that $\gamma \in C_2([0,T])$, that it is strictly positive for $t \in [0,T]$, and that

$$\begin{aligned} \sup_{s\in[0,T]} |Y(s)/n - \gamma(s)| &= o_P\left\{(\log n)^{-1}\right\}, \\ \sup_{s,t\in[0,T],|t-s|\leq C_K b} |\{Y(t)-Y(s)\}/n - \{\gamma(t)-\gamma(s)\}| &= o_P\left\{(nb\log n)^{-1/2}\right\}, \end{aligned}$$

where the constant C_K is defined in assumption A1.

Assumption A2'. Same conditions as in assumption A2 but replacing $o_P\left\{(nb\log n)^{-1/2}\right\}$ with $o_P\left\{(nb)^{-1}\right\}$

Assumption 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 $d > 0$.

Assumption A3'. Same conditions as in assumption A3 but $\alpha \in C_4([0,T])$ and its fourth derivative is Hölder continuous.

Appendix B. Evaluation of the common variance terms

In Section 5 we compare bandwidth selectors by their asymptotic variances which are of the form $S_2 + S_1\Psi$, where Ψ is a factor that differs among bandwidth selectors, while the terms S_1 and S_2 are common for all of them. For both the local linear and the multiplicative biased corrected hazards, the factor Ψ only depends on the chosen kernels so we have evaluated it for some common choices in Table 1. The terms terms S_1 and S_2 however depend on the hazard function α, the exposure function γ and the weighting function w. Here we evaluate the ratio S_1/S_2 for the two hazard estimators, considering a specific choice for these functions. We consider a hazard function of the form $\alpha(t) = \lambda + c\exp(\beta t)$, where λ, c and β are constants. This hazard specification characterizes the Gompertz-Makeham law of mortality, where the empirical magnitudes for the parameters β and c are about 0.085 and 3×10^{31}, respectively. For the weighting function we consider the case $w(t) \equiv 1$, and for the exposure function the case $\gamma(t) = 1_{\{0\le t\le T\}}$, for $T > 0$.

B.1. Local linear estimator

For the local linear hazard estimators the terms S_1 and S_2 are given by

$$S_1^{\mathrm{LL}} = \frac{1}{25}\frac{R(K)^{-7/5}\int \alpha^2(t)w^2(t)\,dt}{\mu_2(K)^{6/5}\left\{\int \alpha''(t)^2\gamma(t)w(t)\,dt\right\}^{3/5}\left\{\int \alpha(t)w(t)\,dt\right\}^{7/5}},$$

$$S_2^{\mathrm{LL}} = \frac{4}{25}\frac{R(K)^{-2/5}\int \alpha''(t)^2\gamma(t)w^2(t)\alpha(t)\,dt}{\mu_2(K)^{6/5}\left\{\int \alpha(t)w(t)\,dt\right\}^{2/5}\left\{\int \alpha''(t)^2\gamma(t)w(t)\,dt\right\}^{8/5}},$$

The ratio $R^{\mathrm{LL}} = S_1^{\mathrm{LL}}/S_2^{\mathrm{LL}}$, for $\gamma(t) = 1_{\{0\le t\le T\}}$ and $w(t) \equiv 1$, is given by

$$R^{\mathrm{LL}} = \frac{S_1^{\mathrm{LL}}}{S_2^{\mathrm{LL}}} = \frac{1}{4R(K)}\frac{\int_0^T \alpha^2(t)\,dt\int_0^T \alpha''(t)^2\,dt}{\int_0^T \alpha(t)\,dt\int_0^T \alpha''(t)^2\alpha(t)\,dt}$$

For the choice $\alpha(t) = \lambda + c\exp(\beta t)$ the above integrals become

$$\int_0^T \alpha(t)\,dt = \frac{c}{\beta}\left(\exp(\beta T) - 1\right) + \lambda T$$

$$\int_0^T \alpha(t)^2\,dt = \frac{c^2}{2\beta}\left(\exp(2\beta T) - 1\right) + \lambda^2 T + \frac{2\lambda c}{\beta}\left(\exp(\beta T) - 1\right)$$

$$
\begin{aligned}
\int_0^T \alpha''(t)^2 \, dt &= \frac{c^2\beta^3}{2}\left(\exp(2\beta T)-1\right) \\
\int_0^T \alpha''(t)^2\alpha(t) \, dt &= \frac{c^3\beta^3}{3}\left(\exp(3\beta T)-1\right) + \frac{\lambda c^2\beta^3}{2}\left(\exp(2\beta T)-1\right)
\end{aligned}
$$

We substitute these results in the expression of R^{LL} and take limits for $T \to \infty$. We only look at the leading terms in the numerator and the denominator (that is $\exp(4\beta T)$) and we get that $R_{LL} \to 3/(16R(K))$, as $T \to \infty$. For the Epanechnikov kernels the limit is $5/16$. Notice that the limit at zero is $R_{LL} \to (4R(K))^{-1}$, which takes the value $5/12$ for the Epanechnikov kernel.

B.2. Multiplicative bias corrected estimator

For the multiplicative bias corrected estimator the terms are

$$
S_1^{\text{MBC}} = \frac{2^{1/3}}{9^2}\frac{R\left(\Gamma_K\right)^{-5/6}\int \alpha(t)^2 w(t)^2 \, dt}{\mu_2(K)^{4/3}\left\{\int h(t)^2\gamma(t)w(t)\, dt\right\}^{1/3}\left\{\int \alpha(t)w(t)\, dt\right\}^{5/3}},
$$

$$
S_2^{\text{MBC}} = \frac{2^{10/3}}{9^2}\frac{R\left(\Gamma_K\right)^{-2/3}\int h(t)^2\gamma(t)w(t)^2\alpha(t) \, dt}{\mu_2(K)^{4/3}\left\{\int \alpha(t)w(t)\, dt\right\}^{2/3}\left\{\int h(t)^2\gamma(t)w(t)\, dt\right\}^{4/3}},
$$

where $h(t) = \alpha(t)(\alpha''(t)/\alpha(t))''$.

We compute the ratio $R^{\text{MBC}} = S_1^{\text{MBC}}/S_2^{\text{MBC}}$ for the same choice of γ and w as before. It yields to the following expression

$$
R^{\text{MBC}} = \frac{S_1^{\text{MBC}}}{S_2^{\text{MBC}}} = \frac{1}{8R(\Gamma_K)^{1/6}}\frac{\int_0^T \alpha^2(t)\, dt \int_0^T h(t)^2 \, dt}{\int_0^T \alpha(t)\, dt \int_0^T h(t)^2\alpha(t)\, dt}
$$

For the choice $\alpha(t) = \lambda + c\exp(\beta t)$ the calculations are as follows:

$$
\begin{aligned}
\int_0^T \alpha(t)\, dt &= \frac{c}{\beta}\left(\exp(\beta T)-1\right) + \lambda T \\
\int_0^T \alpha(t)^2\, dt &= \frac{c^2}{2\beta}\left(\exp(2\beta T)-1\right) + \lambda^2 T + \frac{2\lambda c}{\beta}\left(\exp(\beta T)-1\right)
\end{aligned}
$$

$$
\begin{aligned}
\int_0^T h^2(t)dt &= \int_{\lambda+c}^{\lambda+c\exp(\beta T)} c\beta^7\lambda^2\frac{e^{\beta t}}{y^4}(2\lambda-y)^2 dy = \int_{\lambda+c}^{\lambda+c\exp(\beta T)} \beta^7\lambda\frac{y-\lambda}{y^4}(2\lambda-y)^2 dy \\
&= \int_{\lambda+c}^{\lambda+c\exp(\beta T)} \beta^7\lambda^2\left(\frac{8}{y^3}\lambda^2 - \frac{4}{y^4}\lambda^3 - \frac{5}{y^2}\lambda + \frac{1}{y}\right)dy \\
&= \beta^7\lambda^2\left[\frac{4\lambda^3}{3y^3} - \frac{4\lambda^2}{y^2} + 5\frac{\lambda}{y} + \ln y\right]_{y=\lambda+c}^{y=\lambda+c\exp(\beta T)}.
\end{aligned}
$$

Here

$$h(t) \quad = \quad \alpha(t)\left(\frac{d^2}{dt^2}\left(\alpha''(t)/\alpha(t)\right)\right) = c\beta^4\lambda\frac{e^{t\beta}}{\left(\lambda+ce^{t\beta}\right)^2}\left(\lambda - ce^{t\beta}\right),$$

and we have made the change of variable $y = \alpha(t)$, $dy = c\beta e^{\beta t}dt = \beta(y-\lambda)\ dt$. Similarly

$$\begin{aligned}
\int_0^T h^2(t)\alpha(t)\ dt &= \int_{\lambda+c}^{\lambda+c\exp(\beta T)} c\beta^7\lambda^2\frac{e^{\beta t}}{y^3}(2\lambda-y)^2dy = \int_{\lambda+c}^{\lambda+c\exp(\beta T)}\beta^7\lambda\frac{y-\lambda}{y^3}(2\lambda-y)^2dy \\
&= \int_{\lambda+c}^{\lambda+c\exp(\beta T)}\beta^7\lambda^2\left(\frac{8}{y^2}\lambda^2 - \frac{4}{y^3}\lambda^3 - \frac{5}{y}\lambda + 1\right)dy \\
&= \beta^7\lambda^2\left[\frac{2\lambda^3}{y^2} - \frac{8\lambda^2}{y} - 5\lambda\ln(y) + y\right]_{y=\lambda+c}^{y=\lambda+c\exp(\beta T)}
\end{aligned}$$

We then substitute the above results in the expression of R^{MBC} and take limits for $T \to \infty$. To this goal we only look at the leading terms in the numerator and the denominator and we get that $R_{MBC} \to \infty$ as $T \to \infty$. And the ratio increases to ∞ as $\log(\lambda + c\exp(\beta T))$, this is, at the linear rate βT. The limit for $T \to 0$ is $(8R(\Gamma_K))^{-1/6}$, which for the Epanechnikov kernel is about 0.13.

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.

Gámiz, M. L., Mammen, E., Martínez-Miranda, M. D. and Nielsen, J. P. (2016). Double one-sided cross-validation of local linear hazards. *J. Royal Statist. Soc. B*, **78**, 755–779.

Gámiz, M. L., Mammen, E., Martínez-Miranda, M. D. and Nielsen, J. P. (2017). `DOvalidation`: *Kernel Hazard Estimation with Best One-Sided and Double One-Sided Cross-Validation.* R package version 1.1.0.

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.

Hall, P. and Johnstone, I. (1992). Empirical Functionals and Efficient Smoothing Parameter Selection. *J. Royal Statist. Soc. B*, **54**(2), 475–530.

Hall, P. and Marron, J. S. (1987). Extent to which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation. *Probab. Theory Rel.*, **74**, 567–581.

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

Hiabu, M., Mammen, E., Martínez-Miranda, M. D. and Nielsen, J. P. (2016) In-sample forecasting with local linear survival densities. *Biometrika*, **103**, 843–859.

Jones, M. C., Linton, O. B. and Nielsen, J. P. (1995). A simple bias reduction method for density estimation. *Biometrika*, **82**, 327–38.

Jones, M. C. and Signorini, D. F. (1997). A comparison of higher-order bias kernel density estimators. *J. Am. Statist. Ass.*, **92**, 1063–1073.

Kapotufe, S. and Verma, N. (2017). Time-Accuracy Tradeoffs in Kernel Prediction: Controlling Prediction Quality. *J. Mach. Learn. Res.*, **18**, 1–29.

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., Sperlich, S. and Verrall, R. (2013). Continuous Chain Ladder: Reformulating and generalizing a classical insurance problem. *Expert Syst. Appl.*, **40**(14), 5588–5603.

Muñoz, I. D. and van der Laan, M. J. (2012). Super Learner Based Conditional Density Estimation with Application to Marginal Structural Models. *Int. J. Biostat.*, **7**(1), article 38.

Nielsen, J. P. (1998). Multiplicative bias correction in kernel hazard estimation. *Scand. J. Statist.*, **25**, 541–553.

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

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

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Royal Statist. Soc. B*, **53**, 683–690.

van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007). Super Learner. Statistical Applications of Genetics and Molecular Biology, **6**(25), article 25.