Local linear density estimation for filtered survival data, with bias correction

Local linear density estimation for filtered survival data, with bias correction

By JENS PERCH NIELSEN

Royal & SunAlliance, Codan, Gammel Kongevej 60, 1790 København V, Denmark

npj@codan.dk

CARSTEN TANGGAARD

Department of Finance, Aarhus School of Business, Fuglesangs Alle 4, DK 8210 Aarhus V, Denmark

cat@asb.dk

and M.C. JONES

Department of Statistics, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom

m.c.jones@open.ac.uk

Summary

A class of local linear kernel density estimators based on weighted least squares kernel estimation is considered within the framework of Aalen's multiplicative intensity model. This model includes the filtered data model that, in turn, allows for truncation and/or censoring in addition to accommodating unusual patterns of exposure as well as occurrence. It is shown that the local linear estimators corresponding to all different weightings have the same pointwise asymptotic properties. However, the weighting previously used in the literature in the i.i.d. case is seen to be far from optimal when it comes to exposure robustness, and a simple alternative weighting is to be preferred. Indeed, this weighting has, effectively, to be well chosen in a 'pilot' estimator of the survival function as well as in the main estimator itself. We also investigate multiplicative and additive bias correction methods within our framework. The multiplicative bias correction method proves to be best in a simulation study comparing the performance of the considered estimators. An example concerning old age mortality demonstrates the importance of the improvements provided.

Some key words: Aalen's multiplicative model; Additive bias correction; Censoring; Counting processes; Exposure robustness; Kernel density estimation; Multiplicative bias correction; Old age mortality.

1. Introduction

The topic of this paper is kernel-based nonparametric density estimation for filtered data. The term 'filtered data' covers the important practical problems of censored and truncated data, and combinations thereof. It also covers further, possibly complicated, patterns of exposure as well as occurrence, this being of particular interest in this article. These forms of 'data contamination' are very common in biostatistical and actuarial studies. We will find it convenient to work in the context of Aalen's (1978) multiplicative intensity model which covers the filtered data model of Andersen, Borgan, Gill and Keiding (1988) as a special case.

Local polynomial modelling and, in particular, its local linear special case are very popular in nonparametric regression (e.g. Fan and Gijbels, 1996, Loader, 1999). Transfer of this methodology to density estimation, even in the i.i.d. case, is not totally straightforward and various versions exist (e.g. Lejeune and Sarda, 1992, Jones, 1993, Fan and Gijbels, 1996, Hjort and Jones, 1996, Loader, 1996, Simonoff, 1996). In this paper, we propose a class of local linear kernel density estimators for filtered data based on one of the two main, and closely related, methods for i.i.d. data, namely that based on weighted least squares kernel estimation.

A particular feature of our proposal is that it involves a weighting scheme over and above the localisation weighting provided by the kernel. However, we end up recommending setting this weighting to unity! If a particular alternative weighting is used, the method reduces in the i.i.d. case to that in Jones (1993, Section 5). We show that the pointwise asymptotic properties of the methodology are independent of the particular weighting chosen. However, we come to the recommendation just mentioned because the alternative weighting proves to be much less robust to volatile exposure patterns than the unit weighting.

We go on to consider two bias correction methods which we introduce in the same local least squares framework. One is a multiplicative bias correction, the other an additive bias correction. The first of these is related to the method introduced in nonparametric regression by Linton and Nielsen (1994)

and transferred to i.i.d. density estimation by Jones, Linton and Nielsen (1995) and to kernel hazard estimation by Nielsen (1998). (The latter paper was based on Aalen's multiplicative model as is this paper.) We then introduce to density estimation an additive bias reduction technique that was introduced in the hazard estimation case by Nielsen and Tanggaard (2001). We conclude, using the results of our fairly extensive simulation study, that the multiplicative bias correction is the best of our variations on local linear estimation for density functions. This is in contradistinction to the hazard estimation case where it was found that the use of an additive bias corrected estimator was to be recommended.

The outline of the paper is as follows. In Section 2 we describe the theoretical background in terms of the counting process formulation, including the important special case of filtered data. Two forms of pilot survival function estimator are also described there. In Section 3, particularly Section 3.1, we consider the basic classes of local constant and local linear estimators. All our estimators involve a weighting function W, the choice of which is initially discussed in Section 3.2. We relate the estimators of this section to existing estimators in Section 3.3. In Section 4, we introduce the multiplicative bias correction method which is a general method that can be applied to any initial estimator, although we utilise the local linear estimator as that initial estimator in practice. Like many bias improvement methods (e.g. Jones and Signorini, 1997), those considered here are open to iteration; a double multiplicative bias corrected estimator is, therefore, also defined in this section. Our additive bias reducing principle is introduced in Section 5, where it is combined with a form of multiplicative bias correction (see also Nielsen and Tanggaard, 2001).

In Section 6, we state the pointwise theoretical properties of our estimators. In Section 7, we go through the results of our simulation study comparing estimators and weightings. After setting up the simulations in Section 7.2, we give results for complete data in Section 7.3 and allude to very similar results obtained for censored data. In these cases, choice of weighting is unimportant. But in Section 7.4 we introduce a complex exposure/occurrence pattern and in that case find the choice of weighting to be crucial. In particular, we demonstrate the importance of what we call the exposure robustness of the unit weighting. Cross-validatory bandwidth selection is briefly described in Section 8. In Section 9, an example taken from the actuarial literature concerning old age mortality demonstrates the importance of the improvements provided. Our results and recommenda-

tions are brought together in the closing Section 10. We consider only the univariate case throughout.

2. Counting process background

Consider first the case where $X_1, \ldots, X_n$ are i.i.d. observations. Let $N_i^{(n)}$ indicate observed failures for X_i i.e. $N_i^{(n)}(t) = I(X_i < t)$, with $I(\cdot)$ denoting the usual indicator function. $\mathbf{N}^{(n)} = (N_1^{(n)}, ..., N_n^{(n)})$ is an n-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_t^{(n)},\ t \in [0, T]$, given below, see Andersen, Borgan, Gill and Keiding (1992, p.60). We model the random intensity process $\lambda^{(n)} = (\lambda_1^{(n)}, ..., \lambda_n^{(n)})$ of $\mathbf{N}^{(n)}$ as

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

without restricting the functional form of the hazard function $\alpha(\cdot)$. Here, $Z_i^{(n)}(t) = I(X_i \geq t)$ is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is at risk. We assume that $\mathcal{F}_t^{(n)} = \sigma\,(\mathbf{N}(s), \mathbf{Z}(s); s \leq t)$ where $Z = (Z_1, Z_2, ..., Z_n)$.

We next follow Andersen *et al.* (1988, p.50) and introduce $C_i(t)$, another predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is at risk; this is the filtering (censoring, truncation) process. Let

$$\overline{N}_i^{(n)}(t) = \int_0^t C_i(y) dN_i^{(n)}(y)$$

be the filtered counting process and introduce the filtered filtration $\overline{\mathcal{F}}_t^{(n)} = \sigma\left(\overline{\mathbf{N}}(s), \mathbf{X}, \mathbf{CZ}(s); s \leq t\right)$. Then the random intensity process $\overline{\lambda}^{(n)} = (\overline{\lambda}_1^{(n)}, ..., \overline{\lambda}_n^{(n)})$ of $\overline{N}^{(n)}$ is

$$\overline{\lambda}_i^{(n)}(t) = \alpha\,(t)\, C_i(t) Z_i^{(n)}(t).$$

Of course, if $C_i = 1$, $i = 1, ..., n$, then we are back in the situation of i.i.d. observations. Other important examples of the filtering process involve censoring and truncation. First, let us say that the stochastic variables are right censored (at either a random or a fixed censoring time) at the time points $(R_1, ..., R_n)$. This corresponds to the filtering processes $C_i = I(t < R_i)$. On the other hand, the filtering processes $C_i = I(t \geq L_i)$ correspond to the stochastic variables being left truncated at $(L_1, ..., L_n)$. What is more, the

general filtering principle clearly allows for repeated left truncation and right censoring for the same individuals. The random intensity process $\overline{\lambda}^{(n)}$ above then allows for possibly complicated, but not necessarily well measured or appreciated, patterns of exposure in the population of interest. Correspondingly, robustness to the existence of complex patterns of exposure, and to the possibly ill-defined nature thereof, called *exposure robustness*, will be a feature of the performance of proposed estimators that will be of central interest later in this article.

In the rest of the article, we consider Aalen's multiplicative model that is even more general in scope and notationally simpler than the more intuitive filtered model considered above. We observe n individuals $i = 1, ..., n$. Let N_i count observed failures for the ith individual in the time interval $[0, 1]$. We assume that N_i is a one-dimensional counting process with respect to an increasing, right continuous, complete filtration $\mathcal{F}_t,\ t \in [0, 1]$, i.e. one that obeys *les conditions habituelles*, see Andersen *et al.* (1992, p.60). We model the random intensity as

$$\lambda_i(t) = \alpha(t)Y_i(t)$$

with no restriction on the functional form of $\alpha(\cdot)$. Again, Y_i is a predictable process taking values in $\{0, 1\}$, indicating (by the value 1) when the ith individual is at risk. We assume that $(N_1, Y_1), ..., (N_n, Y_n)$ are i.i.d. for the n individuals.

Each of the density estimators described in the next three sections involves a pilot estimator of the survival function which will generically be written as $\widehat{S}(t)$. In our simulation work (Section 7), two particular survival estimators will be considered. The first arises from estimating the conditional integrated hazard function $\Lambda(t) = \int_0^t \alpha(s)ds$ by the famous Aalen estimator

$$\widehat{\Lambda}(t) = \int_0^t \left\{Y^{(n)}(s)\right\}^{-1} dN_i(s)$$

where $Y^{(n)}(s) = \sum_{i=1}^n Y_i(s)$. The corresponding estimator for the survival function,

$$\widehat{S}_{KM}(t) = \prod_{s \le t} \left\{1 - d\widehat{\Lambda}(s)\right\},$$

is the even more famous Kaplan-Meier product limit estimator. The second, which is more complicated but turns out to have advantages with respect to robustness to complex filtering patterns, is

$$\widehat{S}_{LLH}(t) = \exp\left\{-\int_0^t \widehat{\alpha}_b(s)ds\right\}$$

where $\widehat{\alpha}_b(s)$ is the local linear hazard function estimator using bandwidth b and the unit weighting, as described in Section 5 of Nielsen and Tanggaard (2001). Note that we do not believe it to be worthwhile to employ a more sophisticated bias-corrected hazard rate estimator at this stage.

3. Local constant and local linear estimators

3·1. Main ideas

Let K be a probability density function symmetric about zero and write $K_b(\cdot) \equiv b^{-1}K(b^{-1}\cdot)$ for any bandwidth b. Let $W(s)$ be an arbitrary weight function and let $q_p(z) = \sum_{i=0}^{p} \theta_i z^i$ denote a polynomial of degree p. Then, we can define a local polynomial kernel density estimator based on the local least squares approach of Nielsen (1998); see also Jones (1993). It is given as $\widehat{f}_{p,W}(t) = \widehat{\theta}_0(t) = \widehat{\theta}_0$ where $\widehat{\theta} = (\widehat{\theta}_0, ..., \widehat{\theta}_p)$ and

$$\begin{aligned}\widehat{\theta} &= \arg\min_\theta \lim_{w\to 0} \sum_{i=1}^{n} \int_0^\infty \left(\frac{1}{w}\int_{s-w}^{s} \widehat{S}(u)dN_i(u) - q_p(t-s)\right)^2 \\ &\qquad\qquad \times K_b(t-s)W(s)Y_i(s)ds \\ &= \arg\min_\theta \sum_{i=1}^{n} \int_0^\infty \left(\widehat{S}(s) \triangle N_i(s) - q_p(t-s)\right)^2 K_b(t-s)W(s)Y_i(s)ds. \quad (1)\end{aligned}$$

Minimisation of the criterion function (1) is well defined because the differentiated criterion function is well defined via adoption of the notation $\int \triangle N_i(s)W(s)ds \equiv \int W(s)dN_i(s)$. (Implicit here is the fact that the first squared term in the expansion of the squared bracket in (1) is irrelevant.)

This is entirely parallel to the usual methods of local polynomial regression estimation as in e.g. Wand and Jones (1995), Fan and Gijbels (1996). We will not consider polynomials of degree $p \geq 2$ again in this article, concentrating on the local constant and local linear cases, $p = 0$ and $p = 1$, respectively. These can be explicitly written down. First,

$$\widehat{f}_{0,W}(t) = \frac{\sum_{i=1}^{n} \int_0^\infty K_b(t-s)W(s)Y_i(s)\widehat{S}(s)dN_i(s)}{\int_0^\infty K_b(t-s)W(s)Y^{(n)}(s)ds}.$$

Second,

$$\widehat{f}_{1,W}(t) = \sum_{i=1}^{n} \int \overline{K}_{t,b}(t-s)W(s)Y_i(s)\widehat{S}(s)dN_i(s),$$

where

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

and

$$a_j(t) = \int K_b(t-s)(t-s)^j W(s) Y^{(n)}(s) ds.$$

Notice that

$$\int \overline{K}_{t,b}(t-s) W(s) Y^{(n)}(s) ds = 1, \int \overline{K}_{t,b}(t-s)(t-s) W(s) Y^{(n)}(s) ds = 0,$$

$$\int \overline{K}_{t,b}(t-s)(t-s)^2 W(s) Y^{(n)}(s) ds > 0,$$

so that $\overline{K}_{t,b}$ can be interpreted as a second order kernel with respect to the measure μ, where $d\mu(s) = W(s)Y^{(n)}(s)ds$.

For any given W, we would expect the local linear estimator to be preferable to the local constant estimator, and this is confirmed in both theory and simulations later in the paper. The usual reasons apply (Wand and Jones, 1995, Fan and Gijbels, 1996): the asymptotic bias of the local constant estimator is less appealing than that of the local linear estimator, and more importantly, in the presence of known boundaries, the local linear estimator provides good boundary correction relative to the local constant estimator. The pointwise asymptotic properties of these and the estimators to be introduced in Sections 4 and 5 will be collected together in Section 6.

3·2. Choice of weight function

It will turn out that the pointwise asymptotic behaviour of the local linear estimator is independent of the choice of weighting function W. This is not so for the local constant estimator. (This behaviour mimics that of weighting functions in nonparametric regression.)

Two particular choices of weight function strike us as being particular candidates for use as W. The first is simply a unit, or perhaps 'natural', weighting $W(s) \equiv 1$. The second is to take $W(s) = \{1/Y^{(n)}(s)\} I(Y^{(n)}(s) > 0) = W_0(s)$, say; following Nielsen and Tanggaard (2001), we call this the Ramlau-Hansen weighting. This is motivated largely by observing what the local polynomial estimators reduce to in the i.i.d. case for which $\widehat{S}_{KM}(s) = Y^{(n)}(s)/n$. It is the weighting $W_0(s)$ which yields the least squares local polynomial approach of Lejeune and Sarda (1992) and Jones (1993, Section 5), and this in

turn yields the ordinary kernel density estimate for both $\hat{f}_{0,W}$ and $\hat{f}_{1,W}$, the latter with boundary correction. Under unit weighting, and for i.i.d. data, $\hat{f}_{0,W}$ essentially estimates $f \times S$, where S is the survivor function, and then divides by an estimate of S.

It will also turn out that the asymptotic indifference to the choice of weighting function W of the local linear estimator — and indeed that of more sophisticated estimators to follow — translates to practical indifference also in the cases of i.i.d. and censored data; see Section 7.3. This will prove to be far from the case with regard to (complex) exposure robustness, however; see Section 7.4.

3·3. Related estimators

We indicated in the introduction to this paper that kernel weighted least squares is "one of the two main, and closely related, methods for i.i.d. data"; the other methodology we had in mind is the kernel weighted local likelihood approach of Copas (1995), Hjort and Jones (1996), Loader (1996) and Eguchi and Copas (1998). If we think of density estimation as the limiting case of regression of histogram bin heights against histogram bin centres as the histogram binwidth tends to zero, the former arises from normal-based local regression in that context, the latter from Poisson regression (Simonoff, 1996). Our main reason for following the least squares path is explicitness of estimators and hence computational simplicity; we also suspect the answers obtained will not be very different, and indeed the asymptotic theory for the two will be the same. (One could also, at least in principle, mimic the local likelihood case by introducing a factor of $1/f$ or, in practice, an estimate thereof, into W.)

Nonnegativity is assured for the local constant estimator but not for the local linear. An attractive way of ensuring nonnegativity is to fit a log-linear form (by kernel weighted least squares or otherwise) rather than a linear one (Loader, 1996). We have much sympathy with this approach, but again computational expediency has currently won us over.

There already exist other kernel approaches to density estimation for censored data (e.g. Marron and Padgett, 1987). They are local constant-type estimators. One obvious estimator is a kernel smoothing of the Kaplan-Meier estimator; $W(s) = W_i(s) = \{1/Y_i(s)\}I(Y_i(s) > 0)$ is taken there (Mielniczuk, 1986). A competing estimator smooths only uncensored data and divides by a Kaplan-Meier estimate of the censoring distribution.

As alluded to at the end of Section 2, the Aalen hazard estimator and Kaplan-Meier survival estimator correspond to one another. However, the survival estimator arising from a smoothed Aalen estimator differs from that of smoothing the Kaplan-Meier estimator, and this will prove to be important for exposure robustness later. Mathematically, thinking of things via the Aalen estimator is more natural and allows asymptotic properties for density estimation to be inherited directly from the hazard estimation case. This is because the Aalen estimator gives a martingale while subtracting the compensator, but the Kaplan-Meier estimator does not. In fact, one can get to the martingale from the Kaplan-Meier estimators by integration by parts, resulting in one term involving the Aalen estimator (and hence the martingale) and another of lower order, $O(1/\sqrt{n})$.

4. Multiplicative bias correction

A multiplicative bias correction was introduced for kernel density estimation in Jones, Linton and Nielsen (1995). In this section we take essentially the same approach within our local least squares framework. First, introduce an estimator $\widetilde{f}(t)$ which in practice we will take to be $\widetilde{f}(t) = \widehat{f}_{1,W}(t)$. Then, do a second local linear minimisation which is aimed at estimating the multiplicative error $g_M(t) \equiv f(t)/\widetilde{f}(t)$ of the estimator $\widetilde{f}(t)$ by $\widehat{g}(t)$, say. Thus the multiplicative bias correction density estimator will be

$$\widehat{f}_M(t) = \widetilde{f}(t)\widehat{g}_M(t).$$

So, take $\widehat{g}_M(t) = \widehat{\theta}_0$ where $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1)$ and

$$\begin{aligned}
\widehat{\theta} &= \arg\min_\theta \sum_{i=1}^n \int_0^\infty \left(\widehat{S}(s) \triangle N_i(s) - \{\theta_0 - \theta_1(t-s)\}\widetilde{f}(s)\right)^2 \\
&\qquad\qquad \times K_b(t-s)W(s)Y_i(s)ds \\
&= \arg\min_\theta \sum_{i=1}^n \int_0^\infty \left(\widetilde{f}^{-1}(s)\widehat{S}(s) \triangle N_i(s) - \{\theta_0 - \theta_1(t-s)\}\right)^2 \\
&\qquad\qquad \times K_b(t-s)\widetilde{f}^2(s)W(s)Y_i(s)ds \qquad (2)
\end{aligned}$$

Explicitly, we have that

$$\widehat{g}_M(t) = \sum_{i=1}^n \int \overline{K}_{t,b}^M(t-s)\widetilde{f}(s)W(s)Y_i(s)\widehat{S}(s)dN_i(s)$$

where $\overline{K}^M_{t,b}$ is constructed as $\overline{K}_{t,b}$ in Section 3 but with the weighting function in the $a_j(t)$'s multiplied by the factor $\widetilde{f}^2(s)$.

Note that, in the i.i.d. case, if we obtained our preliminary estimator with the $W_0(s)$ weighting defined in Section 3.2 and our second step estimator using $W(s) = W_0(s)$ in the formulation above, then we would arrive essentially at the estimator considered in Jones, Linton and Nielsen (1995). Estimator $\widehat{f}_M$ amounts essentially to running the local linear estimation process twice and could be iterated further (using, initially, $\widehat{f}_M$ as $\widetilde{f}$). This double multiplicative bias corrected estimator, which we shall refer to as $\widehat{f}_{M2}$ is also considered in Sections 6 and 7.

5. Additive bias correction

In this section we adapt to density estimation the additive bias reducing technique of Nielsen and Tanggaard (2001). In contradistinction to Section 4, consider the additive error $g_A(t) \equiv f(t) - \widetilde{f}(t)$ in using $\widetilde{f}(t)$ to estimate $f(t)$. Again, seek to estimate the error term in local linear fashion. That is, $\widehat{g}_A(t) = \widehat{\theta}_0$ where $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1)$ and

$$\widehat{\theta} \;=\; \arg\min_\theta \sum_{i=1}^n \int_0^\infty \left(\widehat{S}(s) \triangle N_i(s) - \widetilde{f}(s) - \{\theta_0 - \theta_1(t-s)\}\right)^2 \times K_b(t-s) W(s) Y_i(s) ds \qquad (3)$$

and

$$\widehat{f}_A(t) = \widetilde{f}(t) + \widehat{g}_A(t).$$

We present this, however, not to pursue it in the form just given, but rather as a precursor to the development of this idea proposed in the remainder of this section, after this paragraph. The reason that we do not advocate this simple additive bias correction is that it is essentially equivalent to earlier attempts at additive bias correction which effectively result in higher order kernel estimation (Jones, 1995). In particular, (3) results in a fully local linear version of the twicing notion of Stuetzle and Mittal (1979) which has equivalent fourth order kernel $2K - K * K$ where $*$ denotes convolution. In our view, fourth order kernels are not a very successful way to try to improved kernel density estimators (Marron and Wand, 1992, Jones and Signorini, 1997).

Now let $\widetilde{g}_A(t)$ be a preliminary estimate of $g_A(t)$ to which we shall return below. We seek to improve the additive bias correction available via $\widetilde{f}(t) + \widetilde{g}_A(t)$ by introducing a local linear term multiplicatively into the additive bias correction term. Specifically, $\widehat{m}(t) = \widehat{\theta}_0$ where $\widehat{\theta} = (\widehat{\theta}_0, \widehat{\theta}_1)$ and

$$\begin{aligned}\widehat{\theta} &= \arg\min \sum_{i=1}^{n} \int_0^\infty \left[\widehat{S}(s) \triangle N_i(s) - \widetilde{f}(s) - \{\theta_0 - \theta_1(t-s)\}\, \widetilde{g}_A(s)\right]^2 \\ &\qquad \times K_b(t-s) W(s) Y_i(s) ds \\ &= \arg\min \sum_{i=1}^{n} \int_0^\infty \left[\widetilde{g}_A^{-1}(s)\left\{\widehat{S}(s) \triangle N_i(s) - \widetilde{f}(s)\right\} - \{\Theta_0 - \Theta_1(t-s)\}\right]^2 \\ &\qquad \times K_b(t-s) \widetilde{g}_A^2(s) W(s) Y_i(s) ds \qquad (4)\end{aligned}$$

and

$$\widehat{f}_{A|M}(t) = \widetilde{f}(t) + \widehat{m}(t)\widetilde{g}_A(t).$$

Explicitly,

$$\widehat{m}(t) = \sum_{i=1}^{n} \int \overline{K}_{t,b}^{A|M}(t-s) \widetilde{g}_A(s) W(s) Y_i(s) \widehat{S}(s) d\widetilde{N}_i^{A|M}(s)$$

where

$$d\widetilde{N}_i^{A|M}(s) = dN_i(s) - \widehat{S}(s)^{-1} \widetilde{f}(s) ds$$

and $\overline{K}_{t,b}^{A|M}$ is constructed as $\overline{K}_{t,b}$ in Section 3 but with the weighting function in the $a_j(t)$'s multiplied by the factor $\widetilde{g}_A^2(s)$.

It remains to specify $\widetilde{g}_A(t)$. We use a simple smoothed bootstrapping procedure (Efron and Tibshirani, 1993). LetΨ_t be the functional of the underlying data that results in the estimator $\widetilde{f}(t) = \widehat{f}_{1,W}(t)$, that is

$$\widetilde{f}(t) = \Psi_t \left\{(N_1,\, Y_1), .., (N_n,\, Y_n)\right\}.$$

Then let $\overline{f}(t)$ be the bootstrapped estimator of $\widetilde{f}(t)$:

$$\overline{f}(t) = \Psi_t \left\{(\widehat{\Lambda}_1,\, Y_1), .., (\widehat{\Lambda}_n,\, Y_n)\right\}$$

where the $\widehat{\Lambda}_i$'s are the integrated local linear estimators of the observed counting processes: $\widehat{\Lambda}_i(t) = \int_0^t \widetilde{f}(s) ds$. Then

$$\widetilde{g}_A(t) = \overline{f}(t) - \widetilde{f}(t).$$

The key to using what is really a rather naive estimate of $g_A(t)$ is that we allow our local linear device to introduce improvement; in this way, we are able to use just a single bandwidth throughout and avoid difficult 'pilot' estimation questions such as, perhaps, choice of a second bandwidth.

By the way, an approach in which the multiplicative bias correction of Section 4 has the primary role, but an additive element is introduced as well, turns out not to work and so will not be considered here.

6. Pointwise asymptotic theory

In this section, we assume that the following general assumptions hold: The functions $\gamma, w \in C_1([0,1])$ are positive in t. $\Delta_{t,b}$ is defined as the local neighbourhood $\Delta_{t,b} = [t - 10b, t + 10b]$ where b is the bandwidth and $b \to 0$, $nb \to \infty$ as $n \to \infty$. The functions γ, w are the local limits of, respectively, the exposure and the weighting function, that is

$$\sup_{s \in \Delta_{t,b}} \left| Y^{(n)}(s)/n - \gamma(s) \right| \to_P 0$$

and

$$\sup_{s \in \Delta_{t,b}} |W(s)/n - w(s)| \to_P 0.$$

Also, $f \in C_6([0,T])$.

The $\sqrt{n}-$consistency of $\widehat{S}$ implies that $\widehat{S}$ can be substituted by S in all the theoretical considerations. The derivation of theoretical properties of the five estimators considered in this section therefore parallels the derivation of the theoretical properties given in Nielsen and Tanggaard (2001). For all the estimators the strategy is to write the error term $\widehat{f}(t) - f(t)$ as a variable part V_t converging in distribution plus a stable part B_t converging in probability. V_t is not exactly the variance and B_t is not exactly the bias, but V_t and B_t are analytically tractable quantities that are asymptotically equivalent to, respectively, the variance and the bias.

In Table 1 below we give the asymptotic properties of these two terms for each of the five estimators on which we concentrate from here on. These are:

$\widehat{f}_0$, the local constant estimator (Section 3);

$\widehat{f}_1$, the local linear estimator (Section 3);

$\widehat{f}_M$, the multiplicatively bias corrected local linear estimator (Section 4);

$\widehat{f}_{M2}$, the iterated multiplicatively bias corrected local linear estimator (Section 4);

$\widehat{f}_{A|M}$, the enhanced additively bias corrected local linear estimator (Section 5).

Results do not depend on the particular choice of W and, in particular, cover $W(s) = W_0(s)$ and $W(s) = 1$. (The W subscript has therefore been removed from the notation for the first two estimators above.)

Write $\kappa_2 = \int_{-1}^{1} v^2 K(v) dv$. The asymptotic biases in Table 1 reflect the facts that $\widehat{f}_0$ and $\widehat{f}_1$ are standard, second order bias, estimators, $\widehat{f}_M$ is a fourth order bias estimator and $\widehat{f}_{M2}$ and $\widehat{f}_{A|M}$ are sixth order bias estimators. Given that the normalising factor for the variance is $(nb)^{1/2}$ in all cases, the resulting optimal bandwidth and optimal mean squared error magnitudes given in Table 1 follow readily. While estimators with fourth order bias are popular and quite promising (Jones and Signorini, 1997), those with sixth order bias have not often been promoted. A concern is that any further (asymptotic) improvements in bias may be small and compensated for in practice by increases in variance.

* * * Table 1 about here * * *

Also given in Table 1 is a 'variance factor'. The asymptotic expression for each variance term is of the form $V_t = g_\ell U_t$ where $U_t = \{nby(t)\}^{-1} f(t) S(t)$. In each case, g_ℓ is a simple function of K: $g_1 = \int K^2(u) du$, $g_2 = \int \Gamma_K^2(u) du$ where $\Gamma_K(u) = 2K - K * K(u)$ and $g_3 = \int \Delta_K^2(u) du$ where $\Delta_K(u) = K + \Gamma_K - K * \Gamma_K$. Here, $*$ denotes convolution.

7. A simulation study

7·1. Preamble

In this section, we conduct a simulation experiment on the performance of the five estimators, $\widehat{f}_0, \widehat{f}_1, \widehat{f}_M, \widehat{f}_{M2}, \widehat{f}_{A|M}$. The simulation study has much in common with that of Nielsen and Tanggaard (2001) which was for the case of hazard rate estimation, but some of it, especially aspects of the extension to exposure robustness in Section 7.4, necessarily goes some way beyond what was previously done for the hazard case.

7·2. Experimental design and numerical issues

Our experiments utilised each of seven different examples of true densities, labelled $f^k, k = 1, \ldots, 7$. These densities are gamma distributions and mixtures of gamma distributions. The density f^1 is the gamma with parameters $\lambda = 1, r = 1$, where $r/\lambda = 1$ is the mean and $r/\lambda^2 = 1$ is the variance. Thus, f^1 is an exponential distribution, the density f^2 has mean 1.5 and variance 1, while the density f^3 has mean 3 and variance 1. Introduce also the gamma density g with mean 6 and variance 1. Then, the mixtures $f^4, \ldots, f^7$ are constructed from f^2, f^3 and g by using weight vectors, w, given by:

$$\begin{aligned} f^4: &\quad w = (1/2, 1/2, 0), & f^5: &\quad w = (1/2, 0, 1/2), \\ f^6: &\quad w = (0, 1/2, 1/2), & f^7: &\quad w = (1/3, 1/3, 1/3). \end{aligned}$$

This set of densities is portrayed in Figure 1.

* * * FIGURE 1 ABOUT HERE * * *

Simulated complete data sets were constructed as follows; consideration of the simulation of contaminated datasets will be delayed until Section 7.4. First, we defined a grid on the interval $[0, T]$ with gridlength $\delta_M = T/M_0$; this grid is $\{t_j : \ t_j = (j-1)\delta_{M_0},\ j = 1, \ldots, M_0\}$. Next, for a sample of n individuals, the number of failures at time t_i, $N(t_i)$, were generated from the binomial distribution Bin $\left(Y^{(n)}(t_i), f^k(t_i)S^k(t_i)\delta_{M_0}\right)$ where S^k is the survivor function associated with f^k and $Y^{(n)}(t_i)$ is the number at risk at time t_i as before. Computation time was highly dependent on M_0. In general, $M_0 = 100$ was as low as we felt comfortable to go while still giving nice results in the sense that any changes in results for larger M_0 can be explained by simulation noise. Note that we envisage discretization to a grid as a computational device for general use in the implementation of our estimators and not just for the purposes of this simulation study.

The simulations were repeated for $n = 100$, $n = 1000$ and $n = 10000$ individuals. As kernel we used

$$K(x) = \frac{3003}{2048}(1 - x^2)^6 I(-1 < x < 1).$$

All true densities in the study have support $[0, \infty)$; however, we restrict ourselves to estimation on the interval $[0, 10]$, so that $T = 10$. Note that $\int_0^{10} f^k(s)ds > 0.999$, $k = 1, \ldots, 7$.

We report the results of two strategies for bandwidth selection. The first method is based on the *best possible bandwidth.* This amounts to finding for each simulated set of data, $r = 1, ..., R$, the best possible bandwidth b_r, in the sense of having smallest error in estimating the true density. The following is our measure of estimation error:

$$\mathrm{err}_r(\widehat{f}^k_{\ell,r}, f^k) = n^{-1} \int_0^T \left[\widehat{f}^k_{\ell,r}(s) - f^k(s)\right]^2 Y^{(n)}(s) ds \tag{6}$$

for $\ell \in \{0, 1, M, M2, A|M\}, k = 1, \ldots, 7$. We also tried an *average best bandwidth* strategy, which amounts to finding the bandwidth, b_0, which minimises

$$\mathrm{avgerr}(\widehat{f}^k_{\ell}, f^k) = \frac{1}{R} \sum_{r=1}^{R} \mathrm{err}_r(\widehat{f}^k_{\ell,r}, f^k).$$

The error-minimising bandwidths are found by a one-dimensional search routine (the `golden` algorithm plus the `mnbrak` algorithm for initially bracketing the minimum. Both algorithms are from the *Numerical Recipes* library (`http://nr.com`). The search was confined to an appropriate interval $[\alpha, \beta]$ e.g. $[\alpha, \beta] = [1/M_0, 10]$ for $n = 100$. Our experiments showed that there were problems with multiple local extrema. In order to improve the search and make sure that a global minimum was reached we started the search at up to 10 different, equidistant, locations in the interval $[\alpha, \beta]$.

Both of these approaches to bandwidth selection utilise the known density f^k, and are thus unavailable in practice. The first remains valuable, however, as a guide to relative performance of the underlying methods separated from bandwidth selection methodology and as a benchmark for data-driven estimators of bandwidth. The second, the average best bandwidth, can be construed as a reasonable approximation to the performance of a good automatic bandwidth selector.

7·3. Results for complete and censored data

The results for best possible bandwidth and average best bandwidth are given in Tables 2 and 3, respectively. Let us concentrate initially on Table 2. The local constant estimator $\widehat{f}_0$ is inferior to all the other estimators in all cases, except for improving slightly on the local linear estimator $\widehat{f}_1$ for f^3 and f^6 for $n = 1000$ and $n = 10000$. For $n = 100$, the multiplicative density estimator $\widehat{f}_M$ is generally an improvement on $\widehat{f}_1$, but in all but one case $\widehat{f}_{M2}$

fails to improve on $\widehat{f}_M$. The performance of $\widehat{f}_{A|M}$ is best of all for two of the three monotone or unimodal densities f^1 and f^3, but is worst (except for $\hat{f}_0$) for the bimodal densities $f^4, ..., f^7$. Similar effects are observed for $n = 1000$ and $n = 10000$: $\widehat{f}_M$ generally beats $\widehat{f}_1$, $\widehat{f}_{M2}$ only occasionally improves on $\widehat{f}_M$ (even for $n = 10000$) and $\widehat{f}_{A|M}$ remains poor for the bimodal densities while its performance on the simplest densities gets a little worse. For the average best bandwidth, which is arguably closer to something achievable in practice, much the same lessons are learned.

* * * Tables 2 and 3 about here * * *

In both Tables 2 and 3, the average performance column shows reasonable closeness between $\widehat{f}_1$, $\widehat{f}_M$, $\widehat{f}_{M2}$ and $\widehat{f}_{A|M}$, but there is a consistent winner in $\widehat{f}_M$ (and consistent losers in $\widehat{f}_1$, at least for $n > 100$, and, a long way behind, $\widehat{f}_0$). $\widehat{f}_{M2}$ improves its standing a little as n increases, but has not in general outperformed $\widehat{f}_M$ even when n is as large as 10000.

The qualitative results are largely similar to those found in Nielsen and Tanggaard (2001) for similar estimators in the hazard rate case, but with two important exceptions. The first is that for hazards the local constant estimator performed best for the more complicated hazards and for $n = 100$; we no longer observe any such saving grace with $\widehat{f}_0$. The second difference between hazard and density estimation cases is that while the additive bias corrected hazard estimator was much more competitive with the multiplicative bias corrected hazard estimator, the pendulum has swung back towards the multiplicative bias corrected estimator in the density case.

The simulation results given above are all for the use of the unit weighting $W(s) \equiv 1$ in the main stage of definition of the estimators (together with $\widehat{S}_{KM}$ as the initial survival function estimator). It turns out that results for estimators employing the Ramlau-Hansen weighting $W(s) = \{1/Y^{(n)}(s)\}$ $I(Y^{(n)}(s) > 0) = W_0(s)$ (together with $\widehat{S}_{KM}$) are very similar too. This observation extends to a simulation study (not shown) of a situation in which data were censored, the censoring distribution being the same as the distribution generating the data and the average proportion of censored values being 50%. The situation is quite different, however, for complex exposure patterns, as shown in the next subsection.

7·4. Results for exposure robustness

As well as complete, truncated and censored data, the notion of filtered data extends to situations involving (possibly complex) patterns of exposure as well as occurrences. An area in which both exposures and occurrences are often carefully monitored is in actuarial studies. However, it is also very often the case that data gatherers, whilst very carefully tracking occurrences, are rather less alert to the exposure pattern of the individuals under study. For an example in the context of Aids, see Fusaro, Nielsen and Scheike (1993). Given the existence of situations in which exposure patterns may be complex but not entirely well recorded, we are interested in the robustness of the performance of versions of the estimators discussed in this paper to such situations.

To investigate this, we consider a very volatile exposure pattern in which exposures/occurrences are as for the complete data case of Section 8.1 except for the (almost complete) suppression of both exposures and occurrences in the intervals $[0.8, 1.0], [1.8, 2.0], ..., [9.8, 10.0]$. However, to mimic less accurate recording of exposures, we actually set the exposure to 1 in these intervals, meaning that a single observation — rather than no observation — continues to be recorded as an exposure across these intervals. Other aspects of our simulations are as in Section 7.2, except that we present results only for the average best bandwidth.

Interest now particularly centres on two choices that can be made for each of the same five estimators $\widehat{f}_0$, $\widehat{f}_1$, $\widehat{f}_M$, $\widehat{f}_{M2}$ and $\widehat{f}_{A|M}$. These are the form of weighting used in the 'main' stage of definition of the smooth estimators, either unit or Ramlau-Hansen, and the pilot estimator of the survival function used, either Kaplan-Meier (which implicitly uses the Ramlau-Hansen weighting) or $\widehat{S}_{LLH}$ (using the unit weighting). Note that, from here on, when we refer simply to the weighting used we mean the 'main stage weighting' while the choice between $\widehat{S}_{KM}$ and $\widehat{S}_{LLH}$ will be referred to as the choice of survival function estimator. Results are given in Tables 4 to 7. Table 4 corresponds to $\widehat{S}_{KM}$ and Ramlau-Hansen weighting and Table 5 to $\widehat{S}_{KM}$ and unit weighting; these are the two versions of the estimators considered for complete and censored data in Section 7.3. In addition, Table 6 corresponds to $\widehat{S}_{LLH}$ and Ramlau-Hansen weighting and Table 7 to $\widehat{S}_{LLH}$ and unit weighting.

* * * Tables 4 to 7 about here * * *

Let us make comparisons between tables. First, for $n = 100$, the two estimators using Ramlau-Hansen weighting (Tables 4 and 6) yield broadly

comparable results except that $\widehat{f}_{A|M}$ performs rather worse when $\widehat{S}_{LLH}$ estimates survival than when $\widehat{S}_{KM}$ does. (It is also noteworthy that $\widehat{f}_{A|M}$ is generally the best of the five estimators for the Ramlau-Hansen weighting when $n = 1000$ but is worst when $n = 10000$). Unit weighting (Tables 5 and 7), however, improves on Ramlau-Hansen weighting in almost all cases when $n = 100$, and often by quite considerable amounts. The combination of unit weighting and $\widehat{S}_{LLH}$ (Table 7) gives generally better performance than the combination of unit weighting and $\widehat{S}_{KM}$ (Table 5). It is interesting to note that improvement is relatively uniform across estimators and true densities except for the application of the more complex estimators ($\widehat{f}_M$, $\widehat{f}_{M2}$ and $\widehat{f}_{A|M}$) to the unimodal densities. This results, for the combination of seven densities in our tables, in a slight preference for $\widehat{f}_1$ (and unit weighting and $\widehat{S}_{LLH}$) when $n = 100$.

Larger sample sizes make for some differences in relative performance, however. First, it is intriguing to see (Tables 4 and 6) that for $n = 1000$ and $n = 10000$ and Ramlau-Hansen weighting, use of $\widehat{S}_{KM}$ asserts itself in preference to $\widehat{S}_{LLH}$. ($\widehat{f}_M$ is by no means the method of choice for Ramlau-Hansen weighting, by the way.) Again, however, unit weighting (Tables 5 and 7) outperforms Ramlau-Hansen weighting (Tables 4 and 6) in almost all cases, with for this weighting and larger sample sizes, $\widehat{S}_{LLH}$ (Table 7) providing clear and unequivocal improvement over $\widehat{S}_{KM}$ (Table 5).

It might now be said that, with unit weighting, LLH survival estimation and medium to large sample sizes, the relative performances of the five types of estimator revert essentially to how they were for medium and large datasets with no censoring or contamination. Compare Tables 7 and 3, remembering not to be dismayed by apparently better levels of performance when data is contaminated: we continued with estimation error defined by (6), and this includes an exposure weighting, so the absolute values of errors in the tables are not comparable. As for simpler datasets, $\widehat{f}_M$ appears to have an edge over the other estimators.

8. Bandwidth selection by cross-validation

In this section we describe a general cross-validation procedure to select the smoothing parameter for any nonparametric smoother, $\widetilde{f}_\theta$, depending on the smoothing parameter $\theta \in \Theta \in R^k$ in the current context. The procedure is the analogue of least-squares cross-validation or the leave-one-out principle

in i.i.d. regression and density estimation applied to survival data based on Aalens multiplicative model. See, for example, Simonoff (1996) and Loader (1999) and for the related kernel hazard estimation case, Ramlau-Hansen (1983), J.P. Nielsen's unpublished 1990 University of California at Berkeley Ph.D. thesis, "Kernel estimation of densities and hazards: a counting process approach", and Andersen et al. (1992). Ideally, one would like to choose the smoothing parameter as the minimiser of

$$Q_0(\theta) = n^{-1} \sum_{i=1}^{n} \int_0^T \left\{ \widetilde{f}_\theta(s) - f(s) \right\}^2 Y_i(s) ds$$

which is equivalent to minimising

$$n^{-1} \left\{ \sum_{i=1}^{n} \int_0^T \left[\widetilde{f}_\theta(s) \right]^2 Y_i(s) ds - 2 \sum_{i=1}^{n} \int_0^T \widetilde{f}_\theta(s) f(s) \, Y_i(s) ds \right\}.$$

Only the second of these two terms depends on the unknown density and therefore must be estimated from data. We suggest as estimator of $Q_0(\theta)$

$$\widehat{Q}_0(\theta) = n^{-1} \left\{ \sum_{i=1}^{n} \int_0^T \left[\widetilde{f}_\theta(s) \right]^2 Y_i(s) ds - 2 \sum_{i=1}^{n} \int_0^T \widetilde{f}^i_\theta(s) \widehat{S}(s) dN_i(s) \right\}$$

where $\widetilde{f}^i_\theta(s)$ is the estimator arising when the data set is changed by setting the stochastic process N_i equal to 0 for all $s \in [0, T]$ and $\widehat{S}(s)$ is the Kaplan-Meier estimator of the survival function. The cross-validation choice of θ is $\arg\min_\theta \widehat{Q}_0(\theta)$.

This least squares cross-validation bandwidth selection method is presented here for use in the example to follow. We developed the approach quite fully, finding that, in simulations, it worked most of the time without yielding really impressive performance. However, we did encounter difficulties in a sizeable minority of cases caused by the well known effect on least squares cross-validation of data discretisation (e.g. Silverman, 1986, Section 3.4.3). For these reasons, we did not include the method in the presentation of our simulation study. Moreover, in the example, but not in the simulations where we used a single bandwidth b throughout, we found it better to utilise bandwidths $\theta = (b, b/2)$, b in the main density smoothing and $b/2$ when smoothly pilot-estimating the survival function. For general practice, something better than least squares cross-validation will be required.

9. Example: old age mortality

Traditionally, actuaries estimate a hazard function and then calculate the corresponding density function while evaluating annuities. We suggest an approach directly estimating the density function using the methodology detailed in this paper. Below we will estimate the densities of lifetimes for men and women of 90 years of age and above. The data are Swedish and are taken from Lindbergson (2001). They were analysed in Lindbergson (2001) and Fledelius, Guillen, Nielsen and Petersen (2004). While these two studies considered the development of mortality over time, we have accumulated all the data from the considered period, 1988-1997, and we estimate the old age mortality distribution corresponding to this period of time. Data are represented as exposure and occurrence grouped at yearly intervals; the exact grouping methodology is described in Lindbergson (2001). The exposure process is consequently extremely volatile. The data are given in Table 8.

* * * Table 8 about here * * *

We use these data to illustrate the considerable differences, which we believe are improvements, between results from using our preferred estimator, which we will now call $\widehat{f}_{M;U;LLH}$, and a basic kernel estimator using Ramlau-Hansen weighting and Kaplan-Meier survival estimator, $\widehat{f}_{0;RH;KM}$, which corresponds to the usual kernel density estimator when the data are independent and identically distributed. We treated data for women and men separately. We successfully used the cross-validatory bandwidth choice described in Section 8. It gave bandwidths of 2.00 for $\widehat{f}_{0;RH;KM}$ and 3.46 for $\widehat{f}_{M;U;LLH}$ for the data on women, and 2.78 for $\widehat{f}_{0;RH;KM}$ and 3.44 for $\widehat{f}_{M;U;LLH}$ for the data on men. Notice that significantly more women than men get above 90 years old. Therefore, for the basic estimator at least, we expect a bigger smoothing bandwidth to be appropriate while estimating the mortality distribution of women than while estimating the male distribution. Bandwidths are also larger, as expected, for the smoother estimator using multiplicative bias correction and a smoothed survival function.

Estimated densities using the two estimators on the two sets of data are shown in Figure 2. There, we actually estimated the densities at the values 90, 91, ..., 111 and then linearly interpolated the answers as well as cutting the estimates off at 106, above which densities are very small. Notice a similar pattern in comparison in each case: densities appear to be erroneously low

for $\widehat{f}_{0;RH;KM}$ relative to $\widehat{f}_{M;U;LLH}$. Correspondingly, notice that the tails are much thinner for the more naive estimator than for our preferred estimator.

* * * Figure 2 about here * * *

Some aggregated probabilities emphasise the differences between $\widehat{f}_{0;RH;KM}$ and $\widehat{f}_{M;U;LLH}$. For women, the estimated probabilities of dying above 90 years of age are 0.903 and 0.997 for $\widehat{f}_{0;RH;KM}$ and $\widehat{f}_{M;U;LLH}$, respectively; the corresponding values for men are 0.924 and 1.028. Notice that these probabilities should be close to one and that our preferred estimator does a better job in this regard. Likewise, the estimated probabilities of dying above 100 years of age for women are 0.023 and 0.044 for $\widehat{f}_{0;RH;KM}$ and $\widehat{f}_{M;U;LLH}$, respectively; the corresponding values for men are 0.009 and 0.022. Our assumption is that $\widehat{f}_{0;RH;KM}$ significantly underestimates the probability of very old age.

10. Conclusions

As a result of our simulation study, we strongly recommend the use of unit weighting and the survival function estimator $\widehat{S}_{LLH}$ because of the exposure robustness of the corresponding estimators. Note that one aspect of this recommendation is to use unit weighting in both the places it appears in the overall definition of estimators. In exposure robustness terms, proper choice of weighting is a first order effect compared with choice between the five versions of kernel density estimator, which is, relatively, a second order effect. Nonetheless, we are also able to recommend use of the multiplicative bias corrected estimator $\widehat{f}_M$ on the grounds that it seems to perform best – if not always by a great deal – in all cases except perhaps small ($n = 100$) sample sizes and contaminated data, and it still has second best performance in that case. Note again that this recommendation differs from the hazard case, where $\widehat{f}_{A|M}$ was preferred (Nielsen and Tanggaard, 2001). Finally, there remains a need to develop better practical bandwidth selectors for the best estimators described in this paper.

References

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701–26.

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1988). Censoring, truncation and filtering in statistical models based on counting process theory. *Contemp. Math.* **80**, 19–59.

Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1992). *Statistical Models Based on Counting Processes*. New York: Springer.

Copas, J.B. (1995). Local likelihood based on kernel censoring. *J. Roy. Statist. Soc.*, Ser. B **57**, 221–35.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* London: Chapman and Hall.

Eguchi, S. & Copas, J.B. (1998). A class of local likelihood methods and near-parametric asymptotics. *J. Roy. Statist. Soc.*, Ser. B **60**, 709–24.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Applications.* London: Chapman and Hall.

Fledelius, P., Guillen, M., Nielsen, J.P. & Petersen, K. (2004). A comparative study of parametric and nonparametric estimators of old age mortality in Sweden. *J. Actuar. Prac.* in press.

Fusaro, R.E., Nielsen, J.P. & Scheike, T.H. (1993). Marker-dependent hazard estimation — an application to Aids. *Statist. Med.* **12**, 843–65.

Hjort, N.L. & Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619–47.

Jones, M.C. (1993). Simple boundary correction for kernel density estimation. *Statist. Comput.* **3**, 135–46.

Jones, M.C. (1995). On higher order kernels. *J. Nonparametric Statist.* **5**, 215–21.

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

Jones, M.C. & Signorini, D.F. (1997). A comparison of higher order bias kernel density estimators. *J. Amer. Statist. Assoc.* **92**, 1063–73.

Lejeune, M. & Sarda, P. (1992). Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.* **14**, 457–71.

Lindbergson, M. (2001). Mortality among the elderly in Sweden 1988–1997. *Scand. Actuar. J.* 79–94.

Linton, O.B. & Nielsen, J.P. (1994). A multiplicative bias reduction method for nonparametric regression. *Statist. Probab. Lett.* **19**, 181–7.

Loader, C.R. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, 1602–18.

Loader, C.R. (1999). *Local Regression and Likelihood.* New York: Springer.

Marron, J.S. & Padgett, W.J. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *Ann. Statist.* **15**, 1520–35.

Marron, J.S. & Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–36.

Mielniczuk, J. (1986). Some asymptotic properties of kernel estimators of a density function in case of censored data. *Ann. Statist.* **14**, 766–73.

Nielsen, J.P. (1998). Multiplicative bias correction in kernel hazard estimation. *Scand. J. Statist.* **25**, 541–53.

Nielsen, J.P. & Tanggaard, C. (2001). Boundary and bias correction in kernel hazard estimation. *Scand. J. Statist.* **28**, 675–98.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11**, 453–66.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* New York: Springer.

Stuetzle, W. & Mittal, Y. (1979). Some comments on the asymptotic behavior of robust smoothers. In *Smoothing Techniques for Curve Estimation,* Ed. T. Gasser & M. Rosenblatt, pp. 191–5. Berlin: Springer.

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

Table 1. *Asymptotic bias, variance factor V_t/U_t, order of magnitude of optimal bandwidth and of optimal MSE for the five estimators*

Estimator	Bias	Variance factor	Optimal b	Optimal MSE
$\widehat{f}_0(t)$	$\frac{1}{2}\kappa_2 b^2 f''(t)$	g_1	$n^{-1/5}$	$n^{-4/5}$
$\widehat{f}_1(t)$	$\frac{1}{2}\kappa_2 b^2 f''(t)$	g_1	$n^{-1/5}$	$n^{-4/5}$
$\widehat{f}_M(t)$	$\frac{1}{4}b^4\kappa_2^2 f(t)\left(f''/f\right)''(t)$	g_2	$n^{-1/9}$	$n^{-8/9}$
$\widehat{f}_{M2}(t)$	$\frac{1}{8}b^6\kappa_2^3 f(t)\left\{\left(f''/f\right)''\right\}''(t)$	g_3	$n^{-1/13}$	$n^{-12/13}$
$\widehat{f}_{A\|M}(t)$	$\frac{1}{8}b^6\kappa_2^3\left[2f^{(vi)}(t)\right.$ $\left.+\kappa_2\left\{f^{(iv)}(t)\right\}^2/f''(t)\right]$	g_2	$n^{-1/13}$	$n^{-12/13}$

Table 2. *Averages over* $r = 1, ..., R = 1000$ *simulation runs of the values of* $1000 \times \mathrm{err}_r(\widehat{f}^k_{\ell,r}, f^k)$. *These are given for* $n = 100, 1000, 10000$, *true densities* $f^1, ..., f^7$ *and the five estimators* $\ell \in \{0, 1, M, M2, A|M\}$, *in the case of the unit-weighting, survival estimator* $\widehat{S}_{KM}$ *and bandwidth selected as best possible. The column headed 'Average' contains the averages of the values in the* $f^1, ..., f^7$ *columns*

	f^1	f^2	f^3	f^4	f^5	f^6	f^7	Average
$n = 100$								
$\widehat{f}_0$	6.60	7.56	3.80	4.76	5.52	2.99	3.65	4.99
$\widehat{f}_1$	4.83	5.56	3.93	3.43	4.23	2.92	2.70	3.94
$\widehat{f}_M$	5.30	5.05	3.38	3.28	3.87	2.71	2.72	3.76
$\widehat{f}_{M2}$	6.39	5.42	3.20	3.51	4.02	2.74	2.88	4.02
$\widehat{f}_{A\|M}$	4.09	5.22	3.05	3.61	4.71	2.99	2.92	3.80
$n = 1000$								
$\widehat{f}_0$	1.31	1.98	1.09	1.24	1.18	0.65	0.93	1.20
$\widehat{f}_1$	0.73	1.59	1.19	0.90	0.83	0.67	0.63	0.94
$\widehat{f}_M$	0.90	1.23	0.95	0.78	0.68	0.53	0.57	0.81
$\widehat{f}_{M2}$	0.98	1.32	0.87	0.82	0.73	0.50	0.60	0.83
$\widehat{f}_{A\|M}$	0.82	1.31	0.78	0.99	0.92	0.55	0.72	0.87
$n = 10000$								
$\widehat{f}_0$	0.35	0.88	0.61	0.41	0.29	0.19	0.22	0.42
$\widehat{f}_1$	0.13	0.84	0.65	0.36	0.24	0.21	0.16	0.37
$\widehat{f}_M$	0.25	0.70	0.57	0.30	0.18	0.17	0.13	0.33
$\widehat{f}_{M2}$	0.28	0.73	0.54	0.31	0.19	0.15	0.13	0.33
$\widehat{f}_{A\|M}$	0.31	0.72	0.49	0.35	0.23	0.16	0.17	0.35

Table 3. *Averages over $r = 1, ..., R = 1000$ simulation runs of the values of $1000 \times \text{err}_r(\widehat{f}^k_{\ell,r}, f^k)$. These are given for $n = 100, 1000, 10000$, true densities $f^1, ..., f^7$ and the five estimators $\ell \in \{0, 1, M, M2, A|M\}$, in the case of unit-weighting, survival estimator $\widehat{S}_{KM}$ and bandwidth selected as average best. The column headed 'Average' contains the averages of the values in the $f^1, ..., f^7$ columns*

	f^1	f^2	f^3	f^4	f^5	f^6	f^7	Average
$n = 100$								
$\widehat{f}_0$	7.61	8.10	4.07	5.26	5.81	3.15	4.03	5.43
$\widehat{f}_1$	5.72	5.98	4.16	3.79	4.47	3.09	2.98	4.31
$\widehat{f}_M$	5.91	5.66	3.68	3.94	4.25	2.90	3.20	4.22
$\widehat{f}_{M2}$	6.92	6.05	3.59	4.25	4.44	2.92	3.40	4.51
$\widehat{f}_{A\|M}$	5.17	6.27	3.35	4.64	5.29	3.29	3.55	4.51
$n = 1000$								
$\widehat{f}_0$	1.48	2.03	1.13	1.28	1.21	0.68	0.96	1.25
$\widehat{f}_1$	0.82	1.65	1.23	0.94	0.87	0.70	0.66	0.98
$\widehat{f}_M$	0.95	1.29	0.99	0.82	0.72	0.56	0.60	0.85
$\widehat{f}_{M2}$	1.00	1.39	0.91	0.87	0.77	0.53	0.64	0.87
$\widehat{f}_{A\|M}$	1.00	1.41	0.81	1.09	1.01	0.57	0.79	0.96
$n = 10000$								
$\widehat{f}_0$	0.38	0.89	0.61	0.41	0.30	0.20	0.22	0.43
$\widehat{f}_1$	0.13	0.85	0.65	0.36	0.24	0.21	0.16	0.37
$\widehat{f}_M$	0.26	0.71	0.58	0.30	0.18	0.17	0.13	0.33
$\widehat{f}_{M2}$	0.28	0.73	0.55	0.31	0.19	0.16	0.14	0.34
$\widehat{f}_{A\|M}$	0.34	0.73	0.50	0.36	0.24	0.16	0.17	0.36

Table 4. *Averages over $r = 1, ..., R = 1000$ simulation runs of the values of $1000 \times \text{err}_r(\widehat{f}^k_{\ell,r}, f^k)$. These are given for $n = 100, 1000, 10000$, true densities $f^1, ..., f^7$, volatile exposure pattern as described in the text and the five estimators $\ell \in \{0, 1, M, M2, A|M\}$, in the case of the Ramlau-Hansen weighting, survival estimator $\widehat{S}_{KM}$ and bandwidth selected as average best. The column headed 'Average' contains the averages of the values in the $f^1, ..., f^7$ columns*

	f^1	f^2	f^3	f^4	f^5	f^6	f^7	Average
$n = 100$								
$\widehat{f}_0$	9.30	13.10	6.76	7.63	7.73	4.43	5.15	7.73
$\widehat{f}_1$	4.95	11.67	6.76	6.47	6.60	4.43	4.82	6.53
$\widehat{f}_M$	5.72	10.43	6.70	6.29	6.53	4.06	4.56	6.33
$\widehat{f}_{M2}$	6.15	11.48	5.96	6.52	6.45	3.82	4.72	6.44
$\widehat{f}_{A\|M}$	5.41	8.62	4.18	6.20	7.14	4.81	4.59	5.85
$n = 1000$								
$\widehat{f}_0$	2.81	3.39	3.08	3.05	2.90	2.02	2.74	2.85
$\widehat{f}_1$	3.07	3.36	3.08	3.03	2.87	2.02	2.71	2.88
$\widehat{f}_M$	3.23	3.56	3.55	3.54	3.43	3.48	3.38	3.45
$\widehat{f}_{M2}$	1.69	3.75	3.81	3.82	3.77	3.83	3.74	3.49
$\widehat{f}_{A\|M}$	0.96	2.88	1.74	2.69	2.77	1.72	2.07	2.12
$n = 10000$								
$\widehat{f}_0$	0.88	1.27	1.17	1.14	1.08	1.09	0.99	1.09
$\widehat{f}_1$	0.88	1.27	1.17	1.14	1.07	1.09	0.99	1.09
$\widehat{f}_M$	0.85	1.40	1.44	1.36	1.24	1.32	1.17	1.25
$\widehat{f}_{M2}$	0.94	1.40	1.43	1.38	1.26	1.35	1.19	1.28
$\widehat{f}_{A\|M}$	0.99	1.41	1.44	1.40	1.30	1.39	1.23	1.31

Table 5. *Averages over $r = 1, ..., R = 1000$ simulation runs of the values of $1000 \times \mathrm{err}_r(\widehat{f}^k_{\ell,r}, f^k)$. These are given for $n = 100, 1000, 10000$, true densities, $f^1, ..., f^7$, volatile exposure pattern as described in the text and the five estimators $\ell \in \{0, 1, M, M2, A|M\}$, in the case of the unit weighting, survival estimator $\widehat{S}_{KM}$ and bandwidth selected as average best. The column headed 'Average' contains the averages of the values in the $f^1, ..., f^7$ columns*

	f^1	f^2	f^3	f^4	f^5	f^6	f^7	Average
$n = 100$								
$\widehat{f}_0$	8.66	9.16	4.02	6.20	6.94	3.75	4.57	6.19
$\widehat{f}_1$	6.99	5.42	3.15	3.76	4.98	3.56	3.44	4.47
$\widehat{f}_M$	6.00	5.80	3.30	4.69	5.12	3.84	3.94	4.67
$\widehat{f}_{M2}$	6.17	6.22	3.40	4.92	5.34	3.94	4.05	4.86
$\widehat{f}_{A\|M}$	5.62	6.93	3.77	5.26	5.78	4.05	4.01	5.06
$n = 1000$								
$\widehat{f}_0$	1.97	2.53	1.70	2.16	2.13	1.50	1.83	1.97
$\widehat{f}_1$	1.43	1.75	1.09	1.44	1.50	1.36	1.32	1.41
$\widehat{f}_M$	0.96	1.60	1.17	1.56	1.47	1.38	1.37	1.36
$\widehat{f}_{M2}$	0.92	1.69	1.25	1.59	1.57	1.39	1.42	1.40
$\widehat{f}_{A\|M}$	1.22	1.91	1.41	1.80	1.80	1.41	1.46	1.57
$n = 10000$								
$\widehat{f}_0$	0.79	1.32	1.23	1.20	1.09	1.07	0.99	1.10
$\widehat{f}_1$	0.61	1.26	0.79	1.02	0.95	0.94	0.87	0.92
$\widehat{f}_M$	0.36	1.03	0.89	1.03	0.94	0.97	0.87	0.87
$\widehat{f}_{M2}$	0.28	1.07	0.96	1.04	0.96	0.98	0.88	0.88
$\widehat{f}_{A\|M}$	0.54	1.17	1.03	1.14	1.04	1.00	0.91	0.98

Table 6. *Averages over $r = 1, ..., R = 1000$ simulation runs of the values of $1000 \times \mathrm{err}_r(\widehat{f}^k_{\ell,r}, f^k)$. These are given for $n = 100, 1000, 10000$, true densities $f^1, ..., f^7$, volatile exposure pattern as described in the text and the five estimators $\ell \in \{0, 1, M, M2, A|M\}$, in the case of the Ramlau-Hansen weighting, survival estimator $\widehat{S}_{LLH}$ and bandwidth selected as average best. The column headed 'Average' contains the averages of the values in the $f^1, ..., f^7$ columns*

	f^1	f^2	f^3	f^4	f^5	f^6	f^7	Average
$n = 100$								
$\widehat{f}_0$	8.45	13.00	6.56	7.29	7.61	4.36	4.88	7.45
$\widehat{f}_1$	4.81	10.74	6.56	6.02	6.34	4.35	4.55	6.20
$\widehat{f}_M$	5.09	9.83	7.11	5.98	6.19	4.05	4.27	6.08
$\widehat{f}_{M2}$	6.03	11.07	7.22	6.27	6.19	3.84	4.43	6.43
$\widehat{f}_{A\|M}$	6.52	11.00	4.59	6.98	7.62	4.81	4.73	6.61
$n = 1000$								
$\widehat{f}_0$	2.96	3.89	3.56	3.40	3.12	2.03	2.88	3.12
$\widehat{f}_1$	0.62	3.85	3.56	3.38	3.07	2.03	2.12	2.66
$\widehat{f}_M$	2.39	4.31	4.24	4.06	3.76	3.81	3.64	3.74
$\widehat{f}_{M2}$	0.76	4.58	4.56	4.40	4.15	4.21	4.04	3.81
$\widehat{f}_{A\|M}$	1.30	4.00	2.21	3.37	3.23	1.80	2.38	2.61
$n = 10000$								
$\widehat{f}_0$	1.15	1.79	1.68	1.50	1.30	1.33	1.15	1.41
$\widehat{f}_1$	1.15	1.78	1.68	1.50	1.29	1.33	1.15	1.41
$\widehat{f}_M$	0.18	2.01	2.06	1.82	1.56	1.68	1.42	1.53
$\widehat{f}_{M2}$	0.52	2.01	2.06	1.82	1.56	1.68	1.42	1.58
$\widehat{f}_{A\|M}$	1.43	2.01	2.06	1.82	1.56	1.68	1.42	1.71

Table 7. *Averages over $r = 1, ..., R = 1000$ simulation runs of the values of $1000 \times \mathrm{err}_r(\widehat{f}^k_{\ell,r}, f^k)$. These are given for $n = 100, 1000, 10000$, true densities $f^1, ..., f^7$, volatile exposure pattern as described in the text and the five estimators $\ell \in \{0, 1, M, M2, A|M\}$, in the case of the unit weighting, survival estimator $\widehat{S}_{LLH}$ and bandwidth selected as average best. The column headed 'Average' contains the averages of the values in the $f^1, ..., f^7$ columns*

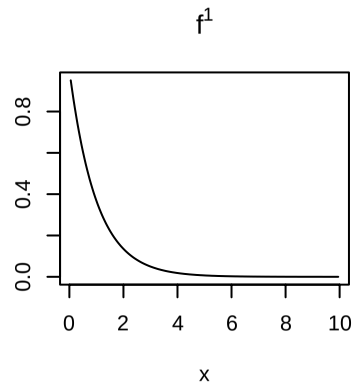
	f^1	f^2	f^3	f^4	f^5	f^6	f^7	Average
$n = 100$								
$\widehat{f}_0$	8.37	7.97	3.90	5.28	6.04	2.90	3.84	5.48
$\widehat{f}_1$	6.37	4.91	3.78	3.32	4.24	2.89	2.77	4.04
$\widehat{f}_M$	5.97	5.04	3.95	3.83	4.44	2.85	3.14	4.17
$\widehat{f}_{M2}$	6.71	5.51	4.26	4.16	4.76	2.96	3.42	4.54
$\widehat{f}_{A\|M}$	5.51	5.81	3.72	4.27	4.94	3.31	3.30	4.42
$n = 1000$								
$\widehat{f}_0$	1.57	1.59	1.02	1.13	1.19	0.66	0.95	1.16
$\widehat{f}_1$	1.00	1.07	0.92	0.68	0.75	0.62	0.57	0.80
$\widehat{f}_M$	0.84	0.74	0.90	0.61	0.66	0.57	0.55	0.69
$\widehat{f}_{M2}$	0.89	0.79	0.98	0.64	0.71	0.59	0.58	0.74
$\widehat{f}_{A\|M}$	0.87	0.90	0.88	0.86	0.92	0.58	0.68	0.81
$n = 10000$								
$\widehat{f}_0$	0.27	0.36	0.24	0.21	0.22	0.17	0.18	0.23
$\widehat{f}_1$	0.15	0.31	0.26	0.19	0.16	0.14	0.12	0.19
$\widehat{f}_M$	0.10	0.20	0.32	0.13	0.12	0.13	0.09	0.15
$\widehat{f}_{M2}$	0.10	0.20	0.32	0.14	0.13	0.13	0.09	0.16
$\widehat{f}_{A\|M}$	0.13	0.25	0.32	0.19	0.17	0.12	0.12	0.19

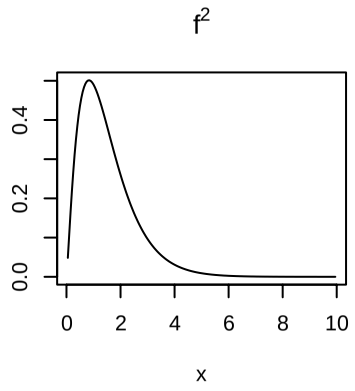
Table 8. *Data on old age mortality in Sweden, 1988-1997*

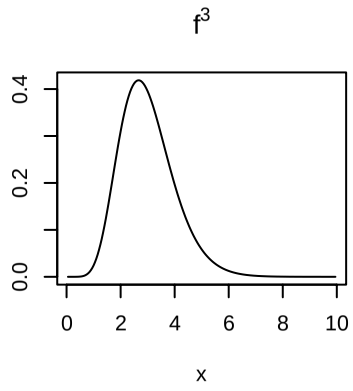
Year	Exposures (Men)	Occurrences (Men)	Exposures (Women)	Occurrences (Women)
90	40191.0	9034	98281.0	16566
91	30370.0	7564	78605.5	15032
92	22477.5	6220	61252.5	13307
93	16155.0	4817	46634.5	10897
94	11389.5	3625	34537.0	9020
95	7847.0	2737	24897.0	7233
96	5202.5	2033	17393.0	5533
97	3366.5	1375	11805.0	4189
98	2095.5	995	7789.5	2841
99	1262.5	573	5070.0	2037
100	753.5	380	3151.0	1433
101	433.5	230	1855.5	896
102	246.0	137	1078.0	534
103	143.0	74	607.0	321
104	76.5	50	334.5	191
105	39.0	21	182.5	90
106	16.5	13	102.5	49
107	8.0	3	49.0	31
108	4.0	4	26.0	16
109	1.5	0	13.0	8
110	0.5	1	3.0	5
111	0.0	0	1.0	0

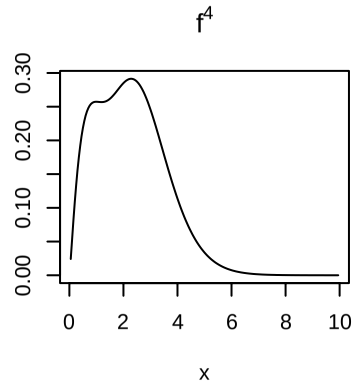
[p.33] Fig. 1. The seven test densities $f^1, ..., f^7$ described at the start of Section 7.1.

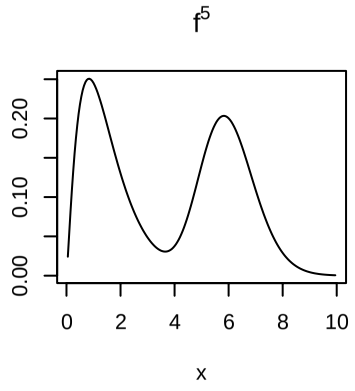
[p.34] Fig. 2. 'Naive' and preferred estimators for old age mortality data on women and men, using cross-validatory bandwidth selection. For women: $\widehat{f}_{M;U;LLH}$ is solid line, $\widehat{f}_{0;RH;KM}$ is dashed line. For men: $\widehat{f}_{M;U;LLH}$ is dot-dashed line, $\widehat{f}_{0;RH;KM}$ is dotted line.

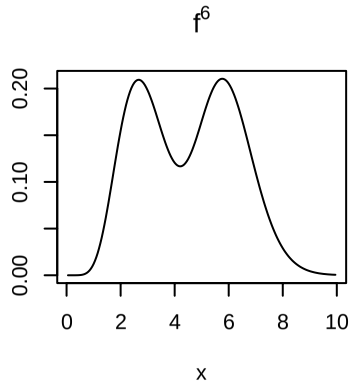

f1
0.0
0.4
0.8
0
2
4
6
8
10
x

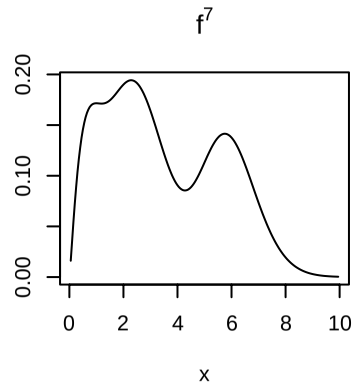

f2
0.0
0.2
0.4
0
2
4
6
8
10
x

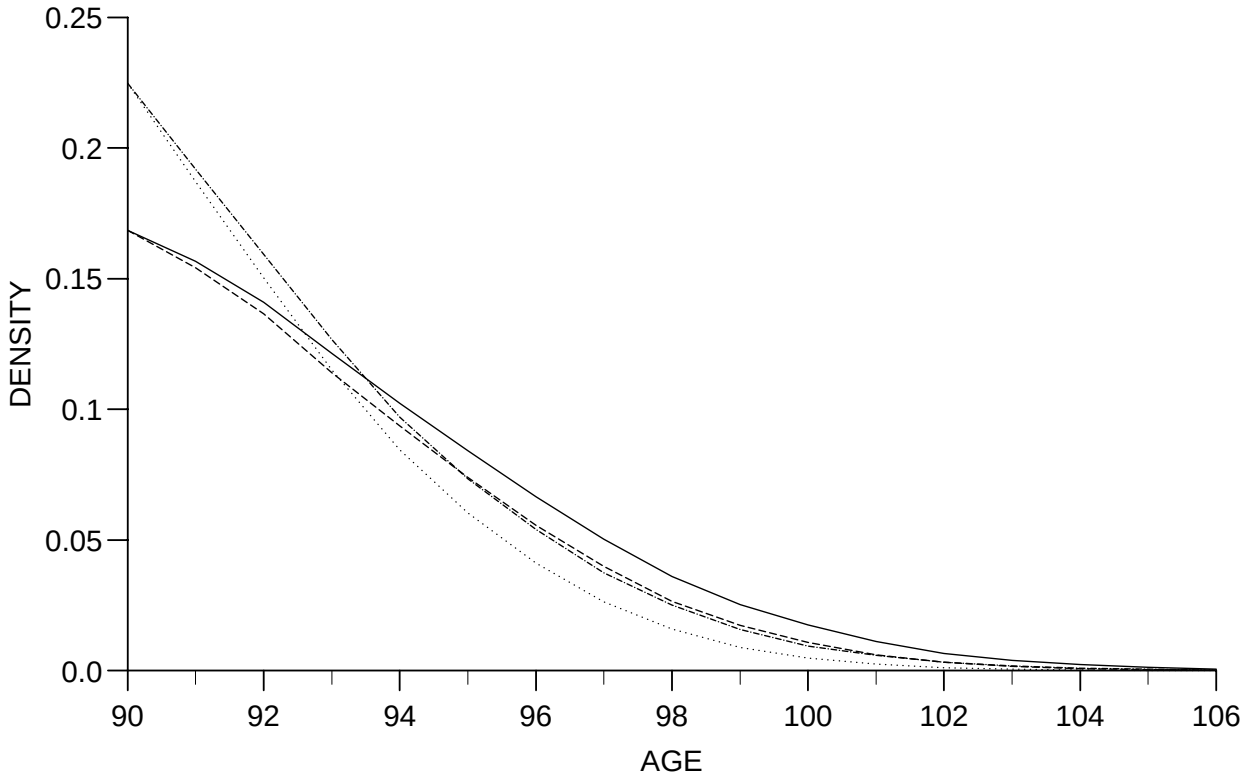

f3
0.0
0.2
0.4
0
2
4
6
8
10
x


f4
0.00
0.10
0.20
0.30
0
2
4
6
8
10
x


f5
0.00
0.10
0.20
0
2
4
6
8
10
x


f6
0.00
0.10
0.20
0
2
4
6
8
10
x


f7
0.00
0.10
0.20
0
2
4
6
8
10
x


0.25
0.2
0.15
0.1
0.05
0.0
DENSITY
90
92
94
96
98
100
102
104
106
AGE