

Smoothing spline density estimation from doubly truncated data

David Bamio¹ and Jacobo de Uña-Álvarez²

Abstract

In astronomy, survival analysis and epidemiology, among many other fields, doubly truncated data often appear. Double truncation generally induces a sampling bias, so ordinary estimators may be inconsistent. In this paper, smoothing spline density estimation from doubly truncated data is investigated. For this purpose, an appropriate correction of the penalized likelihood that accounts for the sampling bias is introduced. The theoretical properties of the estimator are discussed, and its practical performance compared to kernel smoothing is evaluated through simulations. Applications to density estimation for quasar luminosities and age at onset of Parkinson's disease are included. In both the simulation study and the real data illustrations, the smoothing spline estimator outperforms the kernel density estimator due to its smaller variance and robustness against boundary effects.

MSC: 62G07, 62N02.

Keywords: Interval sampling, random truncation, smoothing methods, survival analysis.

1. Introduction

In astronomy, survival analysis and epidemiology, among many other fields, random double truncation may appear. A target variable X is doubly truncated when there exist two random limits (U, V) such that X is observed only when $U \leq X \leq V$. Efron and Petrosian (1999) found double truncation in the analysis of quasar luminosities. Epidemiological studies on event times, on their turn, often proceed through interval sampling. This means that the epidemiological registry is comprised of individuals who underwent the event of interest between two specific calendar times. Interval sampling results in random double truncation, which generally induces biased sampling. See de Uña-Álvarez, Moreira and Crujeiras (2022) for further details and examples.

¹ Instituto Español de Oceanografía (IEO - CSIC), Centro Oceanográfico de Vigo.

² Department of Statistics and Operations Research, Universidade de Vigo, & CITMAga.

Received: February 2026

Accepted: July 2026

In this paper we consider the problem of estimating the density function f of X under double truncation.

Let F denote the cumulative distribution attached to f . Moreira and de Uña-Álvarez (2012) introduced kernel density estimation for doubly truncated data through a convolution of the nonparametric maximum-likelihood estimator (NPMLE) of F in Efron and Petrosian (1999) and a re-scaled kernel function; see Section 2.1 for details. Although this is a natural approach for estimating f , the NPMLE has no explicit form and may suffer from numerical instability and large variance. Also, Xiao and Hudgens (2019) showed that the existence and uniqueness of the NPMLE is not guaranteed and, therefore, the kernel density estimator could not be well-defined. A solution for these NPMLE-related issues is to impose some structure on F before maximizing the likelihood. Methods based on parametric models for the truncating distribution (Moreira and de Uña-Álvarez, 2012) or for the target distribution (Dörre et al., 2021) exist, but such methods involve a preliminary selection of a parametric model. Flexible alternative approaches include smoothing splines, which get free of the NPMLE by considering a penalized likelihood functional (Good and Gaskins, 1971). Although smoothing spline density estimation is well established (Gu, 2013), little is known about its behaviour under random truncation. Importantly, to the best of our knowledge the practical performance of smoothing splines when estimating a doubly truncated density has never been studied in the literature. In this piece of work, we investigate the performance of smoothing spline density estimator based on doubly truncated data, and we compare it to kernel smoothing.

The rest of the paper is organized as follows. In Section 2 the two aforementioned methods to estimate a density function under double truncation are described in detail, and their theoretical properties are reviewed. A simulation study to explore the finite sample performance of the methods is conducted in Section 3 where, in particular, a comparison of smoothing splines and kernel density estimation is performed. Real data illustrations are given in Section 4, while the main conclusions of this piece of work are reported in Section 5. In both the simulation study and the applications to real data the smoothing spline estimator outperforms the kernel density estimator due to its smaller variance and robustness against boundary effects.

2. Methods and theoretical guarantees

2.1. Kernel density estimator

Kernel smoothing is a well-known approach to nonparametrically estimate a smooth density function f from random samples (Wand and Jones, 1985). Let $\{X_i\}_{i=1}^n$ be independent and identically distributed (iid) observations from the target density f . The standard kernel density estimator is defined as the convolution of a re-scaled kernel function $K_h(\cdot) = K(\cdot/h)/h$ and the empirical distribution function of the data, given by $F_n^*(x) = n^{-1} \sum_{i=1}^n I_{[X_i \leq x]}$:

$$f_h(x) = \int K_h(x-t) dF_n^*(t) = \frac{1}{n} \sum_{i=1}^n K_h(x-X_i). \quad (1)$$

In (1), K is a zero-mean density function and $h = h_n$ is a sequence of positive real numbers (called bandwidths, or smoothing parameters) with $h_n \rightarrow 0$ as $n \rightarrow \infty$.

A kernel density estimator is conceptually simple and has a closed form. The trade-off between bias and variance in (1) is automatically controlled through the selection of an appropriate bandwidth parameter h . The estimator (1) can be easily extended to sampling scenarios different from random sampling; for this, one just replaces F_n^* in (1) with a proper substitute.

Let $\{(U_i, V_i, X_i)\}_{i=1}^n$ be iid data with the conditional distribution of (U, V, X) given $U \leq X \leq V$. It is assumed that the truncating couple (U, V) is independent of X . Moreira and de Uña-Álvarez (2012) introduced kernel density estimation for doubly truncated data as

$$f_h(x) = \int K_h(x-t) dF_n(t) \quad (2)$$

where F_n denotes the NPMLE of F under random double truncation. This estimator F_n can be obtained as the maximizer of the conditional likelihood of the X_i given the (U_i, V_i) ,

$$\mathcal{L}_n(F) = \prod_{i=1}^n \frac{dF(X_i)}{F(V_i) - F(U_i-)}, \quad (3)$$

on the class of all the cumulative distributions F . See Efron and Petrosian (1999) for seminal discussions. Note that (2) requires F_n , which has no explicit form and must be computed numerically. As mentioned in the Introduction, the existence and uniqueness of F_n is not guaranteed (Xiao and Hudgens, 2019). A possible solution is to impose some structure on F before maximizing (3). This is what smoothing splines do; details are given in the following subsection.

2.2. Smoothing spline density estimator

2.2.1. Random sampling

In the setting with complete data, iid observations $\{X_i\}_{i=1}^n$ from the target density f are available. Let $\mathcal{X}$ denote the domain of f , which is assumed to be bounded for technical reasons. Good and Gaskins (1971) proposed to carry out the estimation of f by minimizing the penalized functional

$$L(f) + \lambda J(f). \quad (4)$$

Here, $L(f)$ is a goodness of fit measure, usually taken as the minus log-likelihood of the data, $L(f) = -\log \mathcal{L}_n(F)$; $J(\eta)$ is a roughness penalty defined by some notion of smoothness for functions in the domain $\mathcal{X}$, and λ is the smoothing parameter that controls the trade-off between the two opposing goals. In practice, we choose J as the total

curvature of η , $J(\eta) = \int_{\mathcal{X}} (\eta'')^2$, which corresponds to the natural cubic spline. While the boundedness of $\mathcal{X}$ and the choice of J are important for the theoretical developments (J must be a square semi-norm inducing a reproducing kernel Hilbert space on a bounded domain, see Gu (2013)), smoothing parameter selection is the key for the success in practice of penalized likelihood density estimation, since it controls the trade-off between estimation bias and variance. In the simulation study and the real data illustrations λ is selected using the cross-validation method based on the Kullback-Leibler distance described by Gu (2013, Section 7.3). Minimization of (4) takes place in the Hilbert space $\mathcal{H} = \{f: J(f) < \infty\}$ and the minimizer is called a *smoothing spline*.

Assuming $f > 0$ on $\mathcal{X}$, one can use the logistic density transform $f(x) = e^{\eta(x)} / \int_{\mathcal{X}} e^{\eta}$ to estimate η instead, which is free of the positivity ($f \geq 0$) and unity ($\int_{\mathcal{X}} f(x) dx = 1$) constraints Leonard (1978). The penalized likelihood becomes

$$-\frac{1}{n} \sum_{i=1}^n \eta(X_i) + \log \int_{\mathcal{X}} e^{\eta(x)} dx + \lambda J(\eta) \quad (5)$$

and one may search for its minimizer in $\mathcal{H}$. However, one must note that η is determined by f only up to a constant; in order to make the logistic transform one-to-one and the first term in (5) strictly convex, Gu and Qiu (1993) proposed to enforce a side condition on the members of $\mathcal{H} = \{\eta: J(\eta) < \infty\}$ of the form $A\eta = 0$, where A is an averaging operator on $\mathcal{X}$; for instance, one may impose that $\int_{\mathcal{X}} \eta = 0$, or $\eta(x_0) = 0$ for a certain $x_0 \in \mathcal{X}$. Knowing that $L(\eta)$ is continuous and strictly convex in $\mathcal{H} = \{\eta: A\eta = 0, J(\eta) < \infty\}$, if $L(\eta)$ has a minimizer (i.e., the maximum likelihood estimate) in the null space $\mathcal{N}_J = \{\eta: A\eta = 0, J(\eta) = 0\}$ then the existence and uniqueness of the minimizer of (5) in $\mathcal{H}$ is guaranteed (Gu and Qiu, 1993, Th. 3.1).

2.2.2. Biased sampling

In some settings an iid sample from the target density f may not be available. In the case of biased sampling, the values of the target density are selected from the population with probabilities depending on the values themselves, so some corrections of the ordinary estimators are to be made in order to avoid a systematic estimation bias. Consider an iid sample $\{X_i\}_{i=1}^n$ on $\mathcal{X}$ from densities proportional to $w_i(x)f(x)$, where $w_i(x) \geq 0$ are known biasing functions. As in Gu (2013, Section 7.6.1), let $\mathcal{T}$ be an index set and $w(t, x)$, with argument $(t, x) \in \mathcal{T} \times \mathcal{X}$, a function such that the set $\{w(t, \cdot), t \in \mathcal{T}\}$ includes all possible biasing functions with $w(t, \cdot) \neq w(t', \cdot)$ for $t \neq t'$. Within this setting, every biasing function w_i can be written as $w(T_i, \cdot)$ for some $T_i \in \mathcal{T}$, and taking T_i as observations from a probability density $m(t)$, $t \in \mathcal{T}$, one can treat the data as the pairs (T_i, X_i) observed from a two-stage sampling process. Assuming $0 < \int_{\mathcal{X}} w(t, x)f(x)dx < \infty$, the sampling conditional densities $w(t, x)f(x) / \int_{\mathcal{X}} w(t, \cdot)f$ of X given t are well defined. Using once again the logistic transform $f(x) = e^{\eta(x)} / \int_{\mathcal{X}} e^{\eta}$, the sampling conditional den-

sity of X given t becomes

$$\frac{w(t, x)f(x)}{\int_{\mathcal{X}} w(t, \cdot)f} = \frac{w(t, x)e^{\eta(x)}}{\int_{\mathcal{X}} w(t, \cdot)e^{\eta}}, \quad x \in \mathcal{X},$$

leading to the penalized likelihood functional

$$-\frac{1}{n} \sum_{i=1}^n \left\{ \eta(X_i) - \log \int_{\mathcal{X}} w(T_i, x)e^{\eta(x)} dx \right\} + \lambda J(\eta). \quad (6)$$

When biased sampling arises from random truncation, the pairs (T, X) are drawn from a joint density $g(t)f(x)$, $(t, x) \in \mathcal{T} \times \mathcal{X}$, but only the data inside the observable region $A \subset \mathcal{T} \times \mathcal{X}$ are recorded, while those in A^c are lost. In this context, the biasing functions are $w(t, x) = I_{[(t, x) \in A]}$ and $m(t) \propto g(t) \int_{\mathcal{X}} I_{[(t, x) \in A]} f(x) dx$. For instance, for the standard model of independent left-truncation (Klein and Moeschberger, 2003), $g(t)$ is the density of the truncating variable and one has $A = \{(t, x) \in \mathcal{T} \times \mathcal{X} : t \leq x\}$.

2.2.3. Double truncation

Smoothing spline density estimation can be adapted to double truncation by extending the biasing function $w(t, x)$ to allow for a two-dimensional truncating argument $t = (u, v)$. To be specific, recall that under double truncation the available data are the triplets $\{(U_i, V_i, X_i)\}_{i=1}^n$, which follow the conditional distribution of (U, V, X) given $U \leq X \leq V$. The penalized likelihood functional (6) holds with $T_i = (U_i, V_i)$ and $w(u, v, x) = I_{[u \leq x \leq v]}$. Hence, a smoothing spline to consistently estimate f under double truncation can be introduced as the minimizer of (6). Note that, in the notation of Section 1, one has $A = \{(u, v, x) \in \mathcal{T} \times \mathcal{X} : u \leq x \leq v\}$, so the index set $\mathcal{T}$ is a subset of the Euclidean two-dimensional space. Importantly, the potential degenerateness of the kernel density estimator (2) under double truncation is overcome for the smoothing spline estimator thanks to the additional structure that is contained in $\mathcal{H}$, and doubtlessly this is an advantage of the latter method. To the best of our knowledge, the practical performance of the smoothing spline density estimator under double truncation has not been investigated yet.

An interesting feature of double truncation is that the sampling bias may be negligible. This occurs whenever the selection probability for a specific X -value x , $G(x) = P(U \leq X \leq V | X = x) = P(U \leq x \leq V)$, is constant along x . With interval sampling, for instance, it holds $U = V - \tau$ for some positive constant τ (the length of the sampling interval) and, hence, $G(x) = P(x \leq V \leq x + \tau)$. Thus, in the particular scenario of interval sampling, the density of the sampled X_i coincides with the target f when V is uniformly distributed on a convenient support interval; see de Uña-Álvarez (2023) for further motivation. In such a case, one may proceed to estimate f by minimizing the standard penalized likelihood (5), which may entail a variance reduction compared to the minimizer of (6). This is explored through simulations in Section 3.

2.3. Theoretical guarantees

Smoothing splines and kernel smoothing require different conditions for the existence and uniqueness of the density estimator. For the smoothing spline density estimator, $L(\eta)$ is required to be a continuous and strictly convex functional, and $J(\eta)$ is a square seminorm with a finite dimensional null space (Gu and Qiu, 1993). The kernel density estimator (2) is based on the NPMLE of the cumulative distribution function F . A necessary and sufficient condition for the existence and uniqueness of the NPMLE was provided by Vardi (1985), and this condition can be checked graphically as shown by Xiao and Hudgens (2019). Let us define a graph $\mathcal{G}$ with n vertices, each of them representing an observation triplet (U_i, V_i, X_i) , such that there is a directed edge from vertex i to vertex j if and only if $X_j \in [U_i, V_i]$. A directed path is a sequence of edges connecting a sequence of distinct vertices with all the edges directed in the same direction. If, for any two vertices $i, j \in \mathcal{G}$, there exists a directed path from i to j and vice versa, then we say that the graph $\mathcal{G}$ is strongly connected. Then, there exists a unique NPMLE of F if and only if $\mathcal{G}$ is strongly connected; besides, if $\mathcal{G}$ is connected (meaning that there exists an undirected path between each pair of vertices) but not strongly connected, then a NPMLE of F does not exist (Xiao and Hudgens, 2019).

Xiao and Hudgens (2019) provide an example of a doubly truncated dataset for which the NPMLE does not exist. This implies in particular that the kernel density estimator (2) can not be computed. However, minimization of (6) is feasible for this dataset. The intuitive explanation is that smoothing spline estimation introduces a structure in the target density (the Hilbert space attached to the penalty term) before optimization is performed.

For kernel density estimation, conditions ensuring convergence to the target were investigated in Moreira and de Uña-Álvarez (2012), see also Efron and Petrosian (1999). Such conditions are similar to those corresponding to random sampling, once the identifiability of f has been ensured; this means that the sampling probability $G(x)$ must be positive along the support of f . The asymptotic bias of the kernel smoother, of order h^2 , is just the one corresponding to random sampling. However, the asymptotic variance, with order $(nh)^{-1}$, depends on G , being relatively larger for x -points with a small sampling probability $G(x)$. See Moreira and de Uña-Álvarez (2012) for details.

The asymptotic properties of smoothing splines from biased data are collected in Gu (2013). In particular, under certain conditions the in-probability order of the Kullback-Leibler distance between the estimator and the target is $n^{-1}\lambda^{-1/r} + \lambda^p$, where r and p depend on the smoothness of f and $J(f)$; see Gu (2013, Section 9.2). Then, the role of λ in smoothing splines is similar to that of h for kernel smoothing: larger values of λ will introduce more bias in order to reduce the variance. For biased sampling, high-level assumptions on the biasing function $w(t, x)$ for the asymptotic results to hold are given in Gu (2013, Section 9.2.5). In the particular setting of double truncation, such assumptions impose the regularity of the conditional density of X_i given $(U_i, V_i) = (u, v)$, which is $f(x|u, v) \propto f(x)I(u \leq x \leq v)$, and of the joint distribution of the truncating couple. Primitive conditions on the variables X and (U, V) that may ensure such high-

level assumptions involve the identifiability of the target density, which is a non-trivial issue under double truncation. Specifically, the sampling probability for each X -value, $P(U \leq x \leq V)$, must be positive on the support of f . Besides, the possibility that two different distributions of (U, V, X) may generate the same observable data must be ruled out. The reader is referred to Escanciano and de Uña-Álvarez (2026), discussion of their equation (4.2), for further details.

3. Simulation study

A series of simulations were conducted in order to test the effectiveness of the smoothing spline approach to density estimation from doubly truncated data. Along with the bias-corrected smoothing spline density estimator that minimizes (6), the ordinary smoothing spline method based on (5) which ignores the potential sampling bias as well as the non-parametric kernel density estimator (KDE) described in Moreira and de Uña-Álvarez (2012) were considered for comparison purposes. Four different settings for the distribution of observations X and left-truncation limits U were considered, which are

- S1. $X \sim \text{Unif}(0, 1)$, $U \sim \text{Unif}(-1/3, 1)$ (no sampling bias).
- S2. $X \sim \text{Unif}(0, 1)$, $[(3/4)(U + 1/3)]^{1/2} \sim \text{Unif}(0, 1)$.
- S3. $X \sim \text{Beta}(3/2, 5)$, $[(3/4)(U + 1/3)]^{1/2} \sim \text{Unif}(0, 1)$.
- S4. $X \sim \text{Norm}(1/2, 1/10)$, $U \sim \text{Beta}(20, 20)$.

Moreover, three different cases were considered for the right-truncation boundary V , obtained as $V = U + \tau$. The first case corresponds to an interval sampling context, where the observation interval length is constant; we take $\tau = 1/3$. In the second and the third cases, some noise is introduced to the interval length as $\tau \sim \text{Unif}(1/3 - 1/20, 1/3 + 1/20)$ and $\tau \sim \text{Unif}(1/3 - 3/20, 1/3 + 3/20)$ respectively. Therefore, twelve different settings were used for the simulation of $M = 500$ trials $\{(U_i, V_i, X_i)\}_{i=1}^n$ of sizes $n = 100$, $n = 200$ and $n = 400$. The chosen number of trials was moderate due to the long computational times attached to the smoothing spline estimator, which is about five times that of the KDE. For each sample, density estimates from the three methods described above were obtained, producing a total of 108 different scenarios.

The ordinary smoothing spline estimates were computed using the `ssden` function from the `gss` R package (Gu, 2014). To compute the corrected smoothing spline estimates the function `ssden` was modified. Specifically, the `bias` argument of the function `ssden` was expanded to allow for three arguments. This is important since, for doubly truncated data, the sampling condition is given by $w(t, x, \tau) = I_{\{x \in [t, t+\tau]\}}$ where $t = U$ is the left-truncating value and $\tau = V - U$ is the length of the truncation interval, which may be random. Smoothing parameter selection was performed using the cross-validation method based on the Kullback-Leibler distance described by (Gu, 2013, Section 7.3), with $\alpha = 1.4$ as suggested by the author; this is the default within `ssden`.

Finally, the total curvature $J(\eta) = \int_{\mathcal{X}} (\eta'')^2$ was used as penalty, corresponding to the natural cubic spline, and the minus log-likelihood of the data $L(f) = -\log \mathcal{L}_n(F)$ was employed as goodness-of-fit measure. For kernel density estimates, the `densityDT` function from the `DTDA` R package (Moreira, de Uña-Álvarez and Crujeiras, 2021) was used, employing the one-step direct plug-in rule for bandwidth selection, which is one of the selectors recommended by Moreira and Van Keilegom (2013). This bandwidth was passed to the function `densityDT` by choosing the option `bw="DPI1"`. All estimates from the three methods were discretized on a grid consisting of 101 equally spaced points in the $[0, 1]$ support.

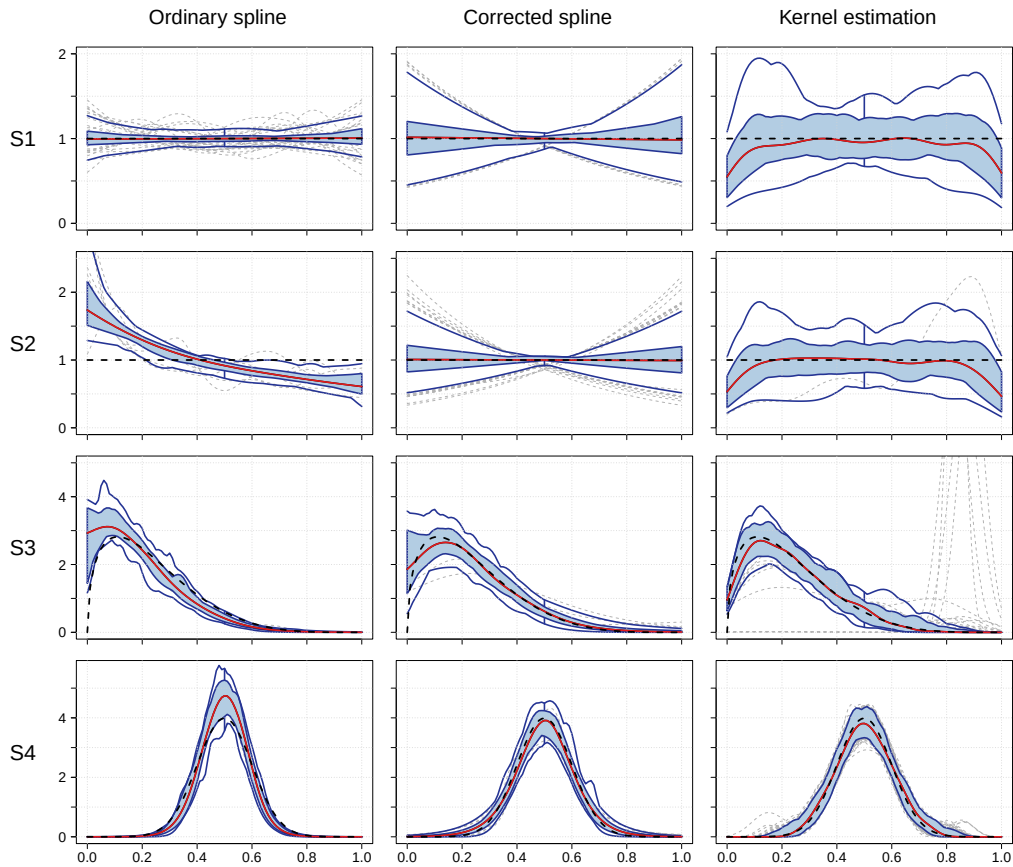


Figure 1. Functional boxplot for 500 Monte Carlo replications of the density estimator with constant width $\tau = 1/3$ of the truncation interval and sample size $n = 400$. Left column: ordinary spline. Middle column: corrected spline. Right column: kernel estimator. The black dashed line corresponds to the target density.

Figures 1, 2 and 3 report functional boxplots for the density estimates obtained for each method and simulation scenario with $n = 400$, allowing for a visual comparison.

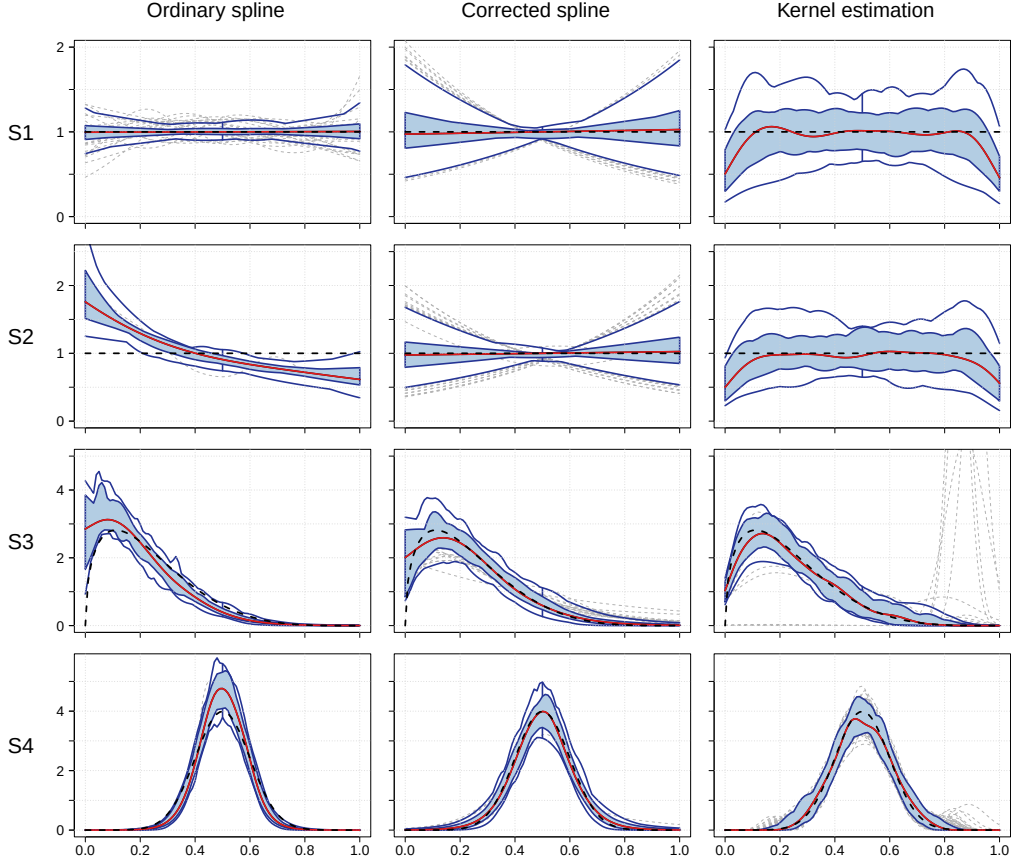


Figure 2. Functional boxplot for 500 Monte Carlo replications of the density estimator with random width $\tau = 1/3 \pm 0.05$ of the truncation interval and sample size $n = 400$. Left column: ordinary spline. Middle column: corrected spline. Right column: kernel estimator. The black dashed line corresponds to the target density.

The black dashed lines indicate the true densities, the red solid lines represent the median curves computed from the 500 estimates, and the light blue bands correspond to the envelope of the deepest 50% of such estimates (López-Pintado and Romo, 2009); finally, the outer blue lines serve to identify outliers. Left columns in these figures reveal that the ordinary smoothing spline estimator is grossly biased. An exception is scenario S1, under which there is no sampling bias. Regarding the S1 and S2 rows, Figures 1 to 3 show that the corrected smoothing spline introduces some noise in the estimation, which increases the variability of the estimates compared to the ordinary smoothing spline. Interestingly, for both S1 and S2 the functional boxplot of the corrected spline presents a bow-tie shape, showing much less variability towards the centre of the support. This is also seen to a small extent for the ordinary spline, and is probably related to the fact

that the uniform density does not vanish at the boundaries. In these two scenarios S1 and S2 the KDE suffers from boundary effects, underestimating the density at points close to 0 or 1. This behaviour may be advantageous for the KDE in scenario S3, where the target density is sharply peaked near the left boundary. In contrast, boundary effects are negligible for S4, since the density vanishes smoothly at both ends of the support in this case.

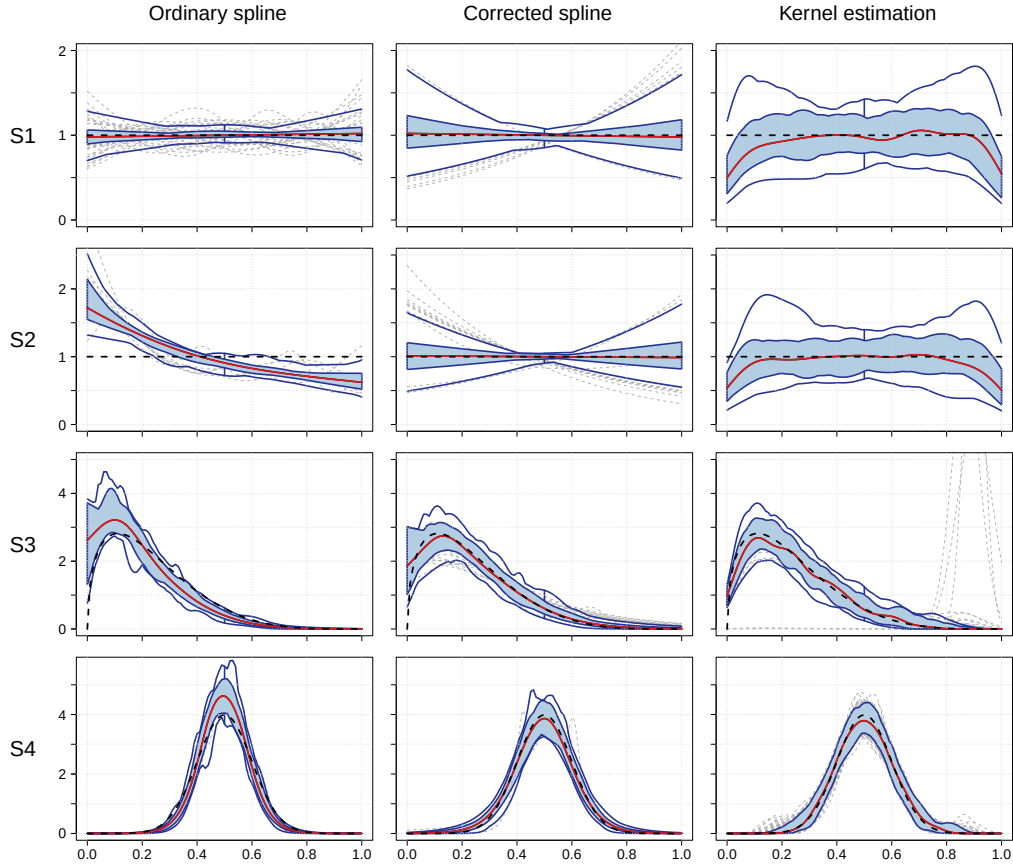


Figure 3. Functional boxplot for 500 Monte Carlo replications of the density estimator with random width $\tau = 1/3 \pm 0.15$ of the truncation interval and sample size $n = 400$. Left column: ordinary spline. Middle column: corrected spline. Right column: kernel estimator. The black dashed line corresponds to the target density.

In order to evaluate the discrepancy between density estimate $\hat{f}$ and the true density f , the integrated squared error $\text{ISE}(f, \hat{f}) = \int_0^1 [f(x) - \hat{f}(x)]^2 dx$ was obtained for each estimate. Table 1 provides the mean (MISE) and standard deviation (SDISE) of the ISE for each simulated scenario. Except for scenario S1 (no sampling bias), the corrected smoothing spline shows significant improvements for $n = 200$ and $n = 400$, where a

systematic estimation bias of the ordinary spline is clearly seen. Moreover, in every case except the S4 scenario, the KDE exhibits a MISE clearly larger than that of the corrected spline estimator. This is partly explained by the fact that some of the weights estimated by KDE are degenerate, leading to a very large value of ISE. This happens in scenario S3 as it can be easily guessed from the large standard deviations attached to the KDE; see also the outliers for the KDE in Figures 1 to 3. For instance, in scenario S3 with $n = 400$ the percentage of degenerate evaluations of the KDE ranges between 13% and 17%. Such degeneration is linked to the aforementioned issues of the NPMLE for doubly truncated data, which may not exist or may not be unique; in such cases, the solution provided by the numerical algorithms may be useless. This instability of the NPMLE under scenario S3 could be related to the fact that the target density is almost vanishing on an interval, $[0.8, 1]$ say, whose points have a relatively small probability of being sampled. This does not occur for scenarios S1, S2 and S4, where the proportion of degenerate NPMLEs (with $n = 400$) is below 3.2%.

To perform a fair comparison that excludes degenerate estimations, Table 2 reports robust measures of the ISE, namely median (MDISE) and interquartile range (IQR-MISE). With this approach, the comparison between the corrected spline and the KDE becomes more ambivalent: the former still performs better for S1 and S2, while KDE shows a small error for S3 and a competitive performance for S4. These results appear to be largely explained by the influence of the boundary effects already observed in Figures 1 to 3. Note that densities in S3 and S4 are well suited to kernel smoothing, since they approach zero at the boundaries, and that the robust summary measures employed in Table 2 mitigate the impact of the outliers for the KDE appearing under scenario S3.

Finally, we investigated the impact of the smoothing parameter in the MISE of the proposed method by considering four additional values ranging from $\alpha = 1$ to $\alpha = 1.8$ in one particular scenario: S2 with $n = 100$. The case with constant $\tau = 1/3$ and one case with random τ (specifically $\tau = 1/3 \pm 0.15$) were considered. See Figure 4. The results were different from those corresponding to the choice $\alpha = 1.4$, with a maximum increase in the MISE of 13% ($\tau = 1/3$, $\alpha = 1.2$), and a maximum decrease of 21% ($\tau = 1/3 \pm 0.15$, $\alpha = 1.6$). This suggests some sensitivity of the method when α varies in a neighbourhood of the default value within `ssden`.

Our simulation results indicate that, under double truncation, a correction of ordinary density estimators which takes the potential sampling bias into account is generally advisable in order to remove the estimation bias. We have illustrated this for the smoothing spline estimator, and we have observed the same pattern for the KDE (results not shown). However, as mentioned, the implementation of such a correction for the KDE may be problematic due to the instability of the NPMLE. For instance, it has been indicated that the KDE degenerates for a number of trials under scenario S3. Although this issue does not affect the smoothing spline estimator, it has been seen how the correction for the sampling bias enlarges the estimation variance, which may be an issue particularly when the sample size is small. Thus, in practice, the application of a preliminary test of negligible sampling bias, such as the one proposed in de Uña-Álvarez (2023), is recommended.

Table 1. Mean ISE, and standard deviation of the ISE in parentheses, of the several density estimators across 500 trials: ordinary smoothing spline (ORD), corrected smoothing spline (COR) and kernel density estimator (KDE). The length of the sampling interval is τ and n is the sample size.

Model		$n = 100$			$n = 200$			$n = 400$		
		ORD	COR	KDE	ORD	COR	KDE	ORD	COR	KDE
$\tau = 1/3 \pm 0.00$	S1	.0126 (.0179)	.1093 (.1546)	.1394 (.1177)	.0067 (.0092)	.0567 (.0762)	.0766 (.0572)	.0033 (.0046)	.0266 (.0345)	.0525 (.0349)
	S2	.1174 (.0775)	.1214 (.1567)	.1407 (.1257)	.1117 (.0532)	.0550 (.0702)	.0875 (.0660)	.1060 (.0402)	.0293 (.0423)	.0542 (.0381)
	S3	.2544 (.1450)	.2372 (.1871)	.3321 (1.6658)	.2037 (.0853)	.1406 (.0996)	.2085 (1.0930)	.1685 (.0569)	.0750 (.0375)	.1144 (.7345)
	S4	.0980 (.0767)	.1010 (.0819)	.0815 (.0638)	.0878 (.0542)	.0512 (.0423)	.0903 (.9162)	.0833 (.0354)	.0246 (.0184)	.0268 (.0189)
$\tau = 1/3 \pm 0.05$	S1	.0139 (.0191)	.0994 (.1329)	.1267 (.1095)	.0055 (.0080)	.0535 (.0696)	.0763 (.0506)	.0035 (.0046)	.0276 (.0373)	.0500 (.0288)
	S2	.1202 (.0768)	.1152 (.1541)	.1416 (.1215)	.1148 (.0558)	.0608 (.0826)	.0872 (.0662)	.1080 (.0390)	.0289 (.0409)	.0499 (.0301)
	S3	.2721 (.1457)	.2453 (.2243)	.3300 (1.5578)	.2194 (.0939)	.1368 (.1109)	.1946 (1.0327)	.1761 (.0579)	.0731 (.0360)	.1104 (.7214)
	S4	.0996 (.0787)	.0972 (.1003)	.0806 (.0647)	.0869 (.0520)	.0527 (.0422)	.0475 (.0304)	.0836 (.0373)	.0258 (.0186)	.0281 (.0205)
$\tau = 1/3 \pm 0.15$	S1	.0143 (.0181)	.0961 (.1176)	.1085 (.0814)	.0066 (.0100)	.0613 (.0802)	.0722 (.0501)	.0039 (.0055)	.0232 (.0326)	.0487 (.0289)
	S2	.1233 (.0774)	.1175 (.1668)	.1245 (.1069)	.1158 (.0574)	.0875 (.1087)	.0805 (.0604)	.1116 (.0390)	.0244 (.0332)	.0518 (.0309)
	S3	.2522 (.1444)	.2208 (.1571)	.2280 (1.0264)	.1940 (.0855)	.1227 (.0779)	.1025 (.5986)	.1628 (.0564)	.0702 (.0301)	.0913 (.6158)
	S4	.0859 (.0734)	.0907 (.0761)	.0770 (.0575)	.0749 (.0453)	.0476 (.0398)	.0477 (.0385)	.0706 (.0316)	.0230 (.0176)	.0267 (.0172)

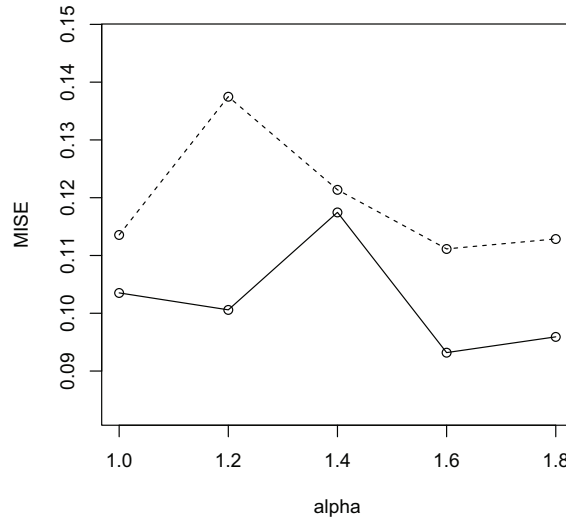


Figure 4. MISE of the smoothing spline estimator for several values of the smoothing factor α , scenario S2 with $n = 100$. Solid line: $\tau = 1/3$. Dashed line: $\tau = 1/3 \pm 0.15$.

Table 2. Median ISE, and interquartile range of the ISE in parentheses, of the several density estimators across 500 trials: ordinary smoothing spline (ORD), corrected smoothing spline (COR) and kernel density estimator (KDE). The length of the sampling interval is τ and n is the sample size.

		$n = 100$			$n = 200$			$n = 400$		
Model		ORD	COR	KDE	ORD	COR	KDE	ORD	COR	KDE
$\tau = 1/3 \pm 0.00$	S1	.0059 (.0162)	.0506 (.1368)	.1024 (.1190)	.0031 (.0079)	.0257 (.0679)	.0593 (.0550)	.0014 (.0047)	.0133 (.0351)	.0416 (.0347)
	S2	.1000 (.0981)	.0563 (.1502)	.1002 (.1207)	.1024 (.0673)	.0260 (.0706)	.0684 (.0605)	.1023 (.0474)	.0130 (.0358)	.0438 (.0347)
	S3	.2177 (.1700)	.1879 (.1895)	.0887 (.0977)	.1905 (.1137)	.1078 (.0882)	.0605 (.0588)	.1634 (.0763)	.0657 (.0382)	.0389 (.0283)
	S4	.0816 (.0884)	.0817 (.1012)	.0658 (.0619)	.0800 (.0649)	.0400 (.0459)	.0398 (.0361)	.0792 (.0438)	.0201 (.0214)	.0223 (.0203)
$\tau = 1/3 \pm 0.05$	S1	.0072 (.0155)	.0520 (.1174)	.0897 (.1108)	.0027 (.0070)	.0292 (.0628)	.0613 (.0585)	.0019 (.0043)	.0129 (.0348)	.0428 (.0300)
	S2	.1079 (.1006)	.0544 (.1459)	.1023 (.1192)	.1069 (.0752)	.0274 (.0786)	.0653 (.0675)	.1042 (.0496)	.0135 (.0353)	.0412 (.0284)
	S3	.2427 (.1921)	.1856 (.1919)	.0933 (.1058)	.2057 (.1331)	.1044 (.0938)	.0612 (.0549)	.1710 (.0753)	.0655 (.0367)	.0388 (.0246)
	S4	.0794 (.0848)	.0662 (.0975)	.0643 (.0700)	.0756 (.0687)	.0421 (.0515)	.0419 (.0396)	.0778 (.0469)	.0204 (.0221)	.0242 (.0199)
$\tau = 1/3 \pm 0.15$	S1	.0070 (.0178)	.0489 (.1257)	.0836 (.0947)	.0031 (.0075)	.0311 (.0737)	.0583 (.0533)	.0019 (.0045)	.0103 (.0261)	.0406 (.0319)
	S2	.1099 (.1027)	.0508 (.1487)	.0927 (.0955)	.1069 (.0750)	.0409 (.1145)	.0614 (.0612)	.1062 (.0524)	.0133 (.0306)	.0424 (.0311)
	S3	.2218 (.1659)	.1697 (.1933)	.0997 (.1016)	.1796 (.1127)	.0989 (.0794)	.0571 (.0530)	.1547 (.0731)	.0641 (.0343)	.0364 (.0270)
	S4	.0657 (.0782)	.0685 (.0905)	.0576 (.0652)	.0674 (.0548)	.0375 (.0424)	.0393 (.0378)	.0675 (.0424)	.0184 (.0183)	.0230 (.0203)

4. Real data illustrations

Two datasets are now introduced to illustrate the practical performance of the smoothing spline density estimator compared to the kernel density estimator. The first dataset contains quasar luminosities, as considered in the seminal paper Efron and Petrosian (1999). The second dataset comprises ages at diagnosis of Parkinson's disease for the late-onset group considered in Clark et al. (2011). These two datasets, which suffer from double truncation, are available within the DTDA R package (Moreira et al., 2021). The density estimates are computed as indicated in Section 3. Functional boxplots are obtained from the simple bootstrap that resamples with replacement from the original dataset $\{(U_i, V_i, X_i)\}_{i=1}^n$; for this, 500 bootstrap replicates are used.

4.1. Quasars data

This example, introduced by Efron and Petrosian (1999) and included in the DTDA package, `Quasars` object, records the luminosity of observed quasars. The original dataset comprises quadruplets (Z_i, m_i, U_i, V_i) , where Z_i is the redshift and m_i is the apparent magnitude of the i th quasar. These two measures are transformed according to the cosmological model assumed (see Efron and Petrosian (1999) for details), obtaining an adjusted luminosity measure $X_i = t(Z_i, m_i)$ in the log-scale. Due to experimental constraints, the distribution of each adjusted log-luminosity X_i is truncated to a known interval $[U_i, V_i]$. Specifically, quasars with apparent magnitude above a certain threshold b_i were too dim to yield dependable redshifts, and the right-truncation limit $V_i = t(Z_i, b_i)$ was obtained as a transformation of b_i . The left-truncation limit $U_i = t(Z_i, a_i)$ was computed from a lower bound a_i on the apparent magnitude, employed in order to avoid confusion with non-quasar astronomical objects. Thus, only those observations with $X_i \in [U_i, V_i]$ are recorded, totalling $n = 210$ triplets (U_i, V_i, X_i) . For this dataset the observation interval length $\tau_i = V_i - U_i$ is not constant.

The results from the three different density estimation methods considered in this paper are displayed in Figure 5 by means of functional boxplots. It can be seen that the bias introduced by the truncation setting is substantial, causing the corrected smoothing spline to yield very different results from the ordinary spline. This sampling bias was formally assessed through the application of the test proposed by de Uña-Álvarez (2023), reporting a p-value of 0; thus, the null hypothesis that the observed data follow the target density was rejected. Although very few quasars at the lower end of the luminosity scale were recorded in the dataset, this is because of the empirical difficulty in observing them; after correcting for bias, most of the probability mass is concentrated at low-luminosity quasars. The kernel density estimates also detect and correct this bias, although the functional interquartile range is wider than that of the smoothing spline estimate. In addition, the kernel density estimator suffers from boundary effects at the left tail. All of these observations are in agreement with the results of the simulation study in Section 3.

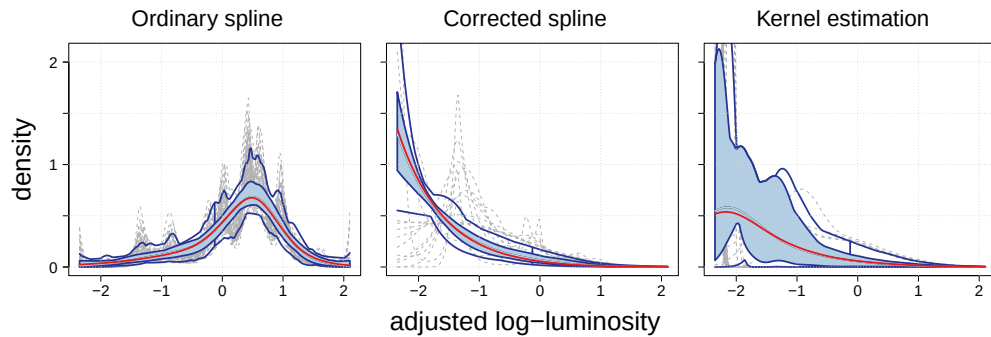


Figure 5. Density estimator computed from *Quasars* data (red line) and functional boxplot for 500 bootstrap replications of the density estimator. Left panel: ordinary smoothing spline. Middle panel: corrected spline. Right panel: kernel estimator.

4.2. Parkinson's disease data

Clark et al. (2011) used these data to investigate the association between two candidate single nucleotide polymorphisms (SNPs) and age of onset of Parkinson's disease (PD). To avoid survival-related bias, the study was restricted to patients whose DNA samples were collected within 8 years of disease onset. As a result, variable X , years of age at onset of PD, is doubly truncated by the age at DNA sampling V , and $U = V - 8$. This leads to an interval sampling setting with a subject-specific sampling interval. The case studies were divided into two groups: an early-onset group, comprising $n = 99$ patients aged 35 to 55 at onset, and a late-onset group, consisting of $n = 100$ patients aged 63 to 87. Here we consider the late-onset group, which is available from the `DTDA` package, `PDlate` object. For this dataset, a formal test rejects the null hypothesis of negligible sampling bias (p-value 0.0022, see de Uña-Álvarez, 2023) and hence a correction of the ordinary density estimators is needed.

Figure 6 displays functional boxplots the density estimates obtained from the several methods. Compared with the ordinary smoothing spline, the bias correction performed by the modified smoothing spline estimator shifts the probability mass to the left. This is because the sampling probability is extremely low for ages below 70 (see de Uña-Álvarez et al., 2022, Example 2.1.8). Interestingly, the corrected smoothing spline estimate exhibits small variance across the bootstrap replicates, in sharp contrast to the large interquartile range of the KDE for ages below 70. As in Figure 5, some boundary effects for the kernel density estimate are also observed at the left tail.

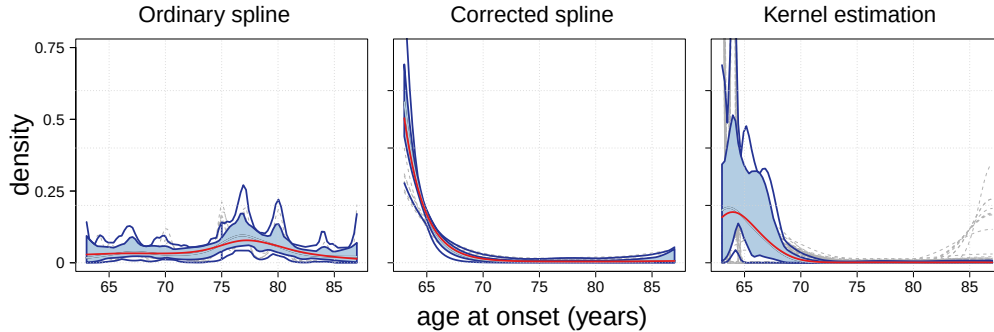


Figure 6. Density estimator computed from `PDlate` data (red line) and functional boxplot for 500 bootstrap replications of the density estimator. Left panel: ordinary smoothing spline. Middle panel: corrected spline. Right panel: kernel estimator.

5. Main conclusions and discussion

In this paper smoothing spline density estimation for doubly truncated data has been investigated. From a theoretical viewpoint, validity of the proposed estimator can be guaranteed from the general theoretical framework corresponding to smoothing splines

with biased data (see Gu, 2013). Practical implementation of the method has been performed through a proper modification of the `ssden` function within the package `gss`. Such modification was needed since `ssden` does not allow for three arguments in the bias function, which is the situation appearing under double truncation. This modification can be found in the repository <https://github.com/DBamio/gss.dtrunc>, where code to run the two real data illustrations in Section 4 is available.

Simulation results have shown that the proposed method performs consistently, approaching the target as the sample size increases. Generally, although not always, the correction for double truncation results in a larger estimation variance; however, it is critical to remove the systematic bias of the ordinary smoothing spline estimator, which may arise from the biased sampling. Compared to other nonparametric estimation strategies such as kernel smoothing, the smoothing spline method is superior in our simulated settings. Two factors are important for this. First, kernel smoothing may suffer from a large variability due to the noise of the NPMLE F_n , which is convolved with the kernel function; moreover, such NPMLE may degenerate due to non-uniqueness or non-existence issues, leading to inadmissible density estimates. This problem does not affect the smoothing spline estimator, which does not use the NPMLE. Second, boundary effects, which are often present in kernel density estimation but naturally avoided by smoothing splines, can lead to a poorer performance of the former approach. These factors have been illustrated and discussed in Sections 3 and 4.

One drawback of smoothing spline density estimation, compared to kernel smoothing, is that it involves a larger computational time. For illustration, for the dataset on quasar luminosities ($n = 210$) the execution time of the adapted `ssden` function was 0.8 seconds, compared with 0.16 seconds for the `densityDT` function. This additional computational effort becomes particularly important when performing bootstrapping for a large number of Monte Carlo trials.

Acknowledgements

Thanks to the Editor and to two anonymous reviewers for comments and suggestions that have served to improve the manuscript. Work supported by the Grant PID2023-148811NB-I00 from Agencia Estatal de Investigación (Ministerio de Ciencia, Innovación y Universidades, Spanish Government). This paper has been presented as invited oral talk at the 26th International Conference on Computational Statistics (COMPSTAT2024) held in Giessen, Germany, August 27-30 2024, and at the 4th IMS International Conference on Statistics and Data Science (ICSIDS2025) held in Seville, Spain, December 15-18 2025; authors acknowledge feedback from the attendees.

References

- Clark, J., Reddy, S., Zheng, K., Betensky, R. A. and Simon, D. K. (2011). Association of PGC-1 α polymorphisms with age of onset and risk of Parkinson's disease. *BMC Medical Genetics*, **12**, 69.
- de Uña-Álvarez, J. (2023). Testing for an ignorable sampling bias under random double truncation. *Statistics in Medicine*, **42**(20), 3732–3744.
- de Uña-Álvarez, J., Moreira, C. and Crujeiras, R. M. (2022). *The Statistical Analysis of Doubly Truncated Data: With Applications in R*. Wiley, Chichester, UK.
- Dörre, A., Huang, C.Y., Tseng, Y.K. and Emura, T. (2021). Likelihood-based analysis of doubly-truncated data under the location-scale and AFT model. *Computational Statistics*, **36**, 375–408.
- Efron, B. and Petrosian, V. (1999). Nonparametric methods for doubly truncated data. *Journal of the American Statistical Association*, **94**, 824–834.
- Escanciano, J.C. and de Uña-Álvarez, J. (2026). Goodness-of-fit tests for censored and truncated data: maximum mean discrepancy over regular functionals. arXiv:2602.08108.
- Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58**(4), 255–277.
- Gu, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. Springer: New York, NY, USA.
- Gu, C. (2014). Smoothing Spline ANOVA Models: R Package `gss`. *Journal of Statistical Software*, **58**(5), 1–25. URL <http://www.jstatsoft.org/v58/i05/>.
- Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: theory. *Annals of Statistics*, **21**(1), 217–234.
- Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data, Second Edition*. Springer, New York.
- Leonard, T. (1978). Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, **40**, 113–132.
- López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, **104**, 486–503.
- Moreira, C. and de Uña-Álvarez, J. (2012). Kernel density estimation with doubly truncated data. *Electronic Journal of Statistics*, **6**, 501–521.
- Moreira, C., de Uña-Álvarez, J. and Crujeiras, R. M. (2021). *DTDA: Doubly Truncated Data Analysis*, R package version 3.0.1. URL <http://cran.r-project.org/package=DTDA>.
- Moreira, C. and Van Keilegom, I. (2013). Bandwidth selection for kernel density estimation with doubly truncated data. *Computational Statistics and Data Analysis*, **61**, 107–123.
- Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, **13**(1), 178–203.

- Wand, M. P. and Jones, M. C. (1985). *Kernel Smoothing. Monographs on Statistics and Applied Probability*. Wiley, London.
- Xiao, J. and Hudgens, M. G. (2019). On nonparametric maximum likelihood estimation with double truncation. *Biometrika*, **106**(4), 989–996.