Supplementary materials for "Kernel distribution estimation for grouped data"

Miguel Reyes[1], Mario Francisco-Fernández[2], Ricardo Cao[2] and Daniel Barreiro-Ures[2]

December 2019

The material contained herein is supplementary to the article named in the title and published in SORT-Statistics and Operations Research Transactions Volume 43(2).

---

[1] Departamento de Actuaría, Física y Matemáticas. Universidad de las Américas-Puebla, Puebla, México. miguel.reyes@udlap.mx

[2] Research Group MODES. Departamento de Matemáticas, Facultade de Informática, CITIC, ITMATI. Universidade da Coruña, A Coruña, Spain. mariofr@udc.es, ricardo.cao@udc.es, daniel.barreiro.ures@udc.es

1 Simulation study

In this section, the simulation study presented in the main paper is completed with some additional experiments. The simulation study shown in the main paper tried to mimic the asymptotic conditions on $\bar{l}$ in Assumption 3.4. Although it can be of interest to analyze situations in which it is ideally observed the sample size increasing and the average length decreasing at different rates, in practice that seldom really occurs. The simulation study presented here deals with a more factual situation in which there is a given sample size and a given set of fixed intervals.

For this simulation, a sample size of $n = 240$, a fixed set of average lengths and a grid of values for h were considered. As the populational density, we used the same normal mixture to that employed in the simulation study presented in the main paper: $f(x) = \sum_{i=1}^{4} \alpha_i \phi_{\mu_i,\sigma_i}$, with $\phi_{\mu,\sigma}$ a $N\left(\mu,\sigma^2\right)$ density, $\alpha = (0.70, 0.22, 0.06, 0.02)$, $\mu = (207, 237, 277, 427)$ and $\sigma = (25, 20, 35, 50)$, where α, μ and σ are the mixture weights, means and standard deviations, respectively.

These were the steps followed:

1. Simulate an n-size sample from the normal mixture reference density f.
2. Divide the data range into intervals such that its average length is $\bar{l}$.
3. Select h and compute MISE_g.
4. Repeat the previous steps considering a grid of possible pairs $\left(\bar{l}, h\right)$.

Figure 1 shows the natural logarithms of MISE_g as function of the average length, $\bar{l}$, and the bandwidth, h, for a fixed sample size of $n = 240$. From the practical viewpoint, it seems interesting the minimum closest to the bottom left, where the estimator reaches its best performance and clearly corresponds to light grouping cases. This area is characterized by an average interval length of $\bar{l} \approx 23$ units or less, or dividing by the average range $\bar{r}$, by a ratio $\bar{l}/\bar{r} \approx 0.08$ or less, which gives a clue about when a good performance of the estimator $\hat{F}_h^g$, given in equation (5) of the main paper, is expected in a practical situation, as long as the sample size is about 240.

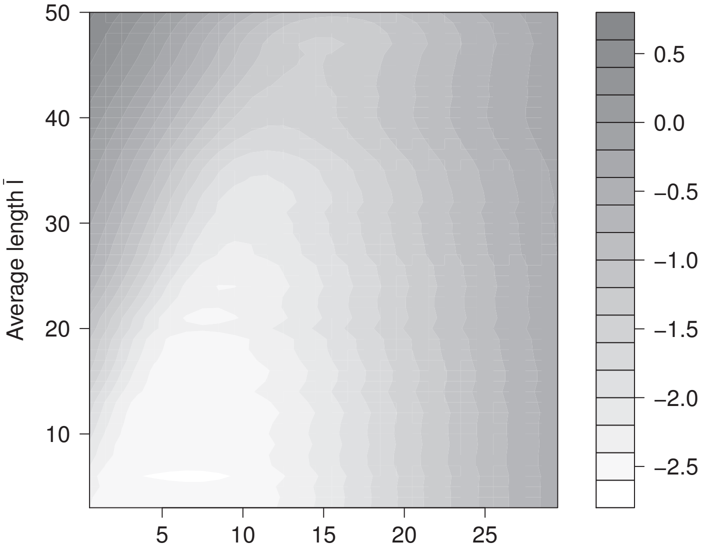

Figure 1: *Natural logarithm of* MISE_g *(gray scale) by bandwidth, h, and average length,* $\bar{l}$*, for a fixed sample size* $n = 240$*.*

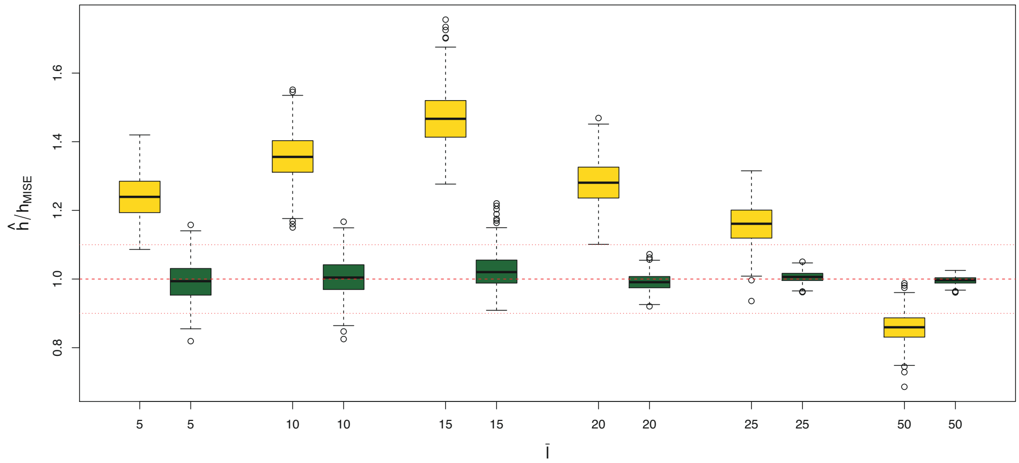

Figure 2: *Sampling distribution of $\hat{h}_{PB_g}/h_{MISE_g}$ (yellow left box-plots for each value of $\bar{l}$) and sampling distribution of h^*_{MISE}/h_{MISE_g} (green right box-plots for each value of $\bar{l}$), for different average lengths and sample size n=240. Red dotted lines are plotted at values 0.9 and 1.1 for reference.*

Figure 2 shows the performance of the plug-in bandwidth, $\hat{h}_{PB_g}$, and the bootstrap selector, h^*_{MISE}, given in equations (15) and (17) of the main paper, respectively, for different values of $\bar{l}$. It is observed that while the sampling distribution of $\hat{h}_{PB_g}$ has, more or less, constant variability for different values of $\bar{l}$, it is also clearly biased. The plug-in selectors are larger than the target value for small or moderate values of $\bar{l}$, and smaller than the optimal bandwidth for large values of $\bar{l}$. On the other hand, the sampling distribution of the bootstrap smoothing parameter, h^*_{MISE}, seems to be very stable, always centered somewhere around h_{MISE_g} and with decreasing variability when the average length $\bar{l}$ increases.

Figure 3 shows a visual example of what may happen when estimating the distribution function for small or large interval average lengths, specifically, for $\bar{l} = 5$ and $\bar{l} = 50$. In the case of $\bar{l} = 5$ (left panel), using the optimal h_{MISE_g} and the estimated bandwidths, $\hat{h}_{PB_g}$ and h^*_{MISE}, the three estimates are indistinguishable. Although $\hat{h}_{PB_g}$ takes a value of 8.87, slightly larger than $h_{MISE_g} = 7.03$ and $h^*_{MISE} = 7.14$, this seems not to have an important influence in the final distribution estimation. Something similar can be observed in the large grouping case with $\bar{l} = 50$ (right panel). In this case, $\hat{h}_{PB_g} = 12.72$ is smaller than the optimal bandwidth, $h_{MISE_g} = 14.71$, and the bootstrap selector, $h^*_{MISE} = 14.82$, but once again the corresponding cdf estimates present a very similar shape. This fact reinforces the argument given at the end of the Simulation Section of the main paper, pointing out that bandwidth selection is not so critical in distribution as in density estimation, since slightly different bandwidths produce very similar distribution estimates.

2 An empirical study from real data

In this section, a real data empirical study was performed considering the time between eruptions of the Old Faithful geyser in the Yellowstone National Park, Wyoming, United States. The data set is available in the `R` environment for statistical computing. It is worth noting that the sample size of this data set is 272, similar to 240, the sample size considered in the previous simulation study, which favors for comparison.

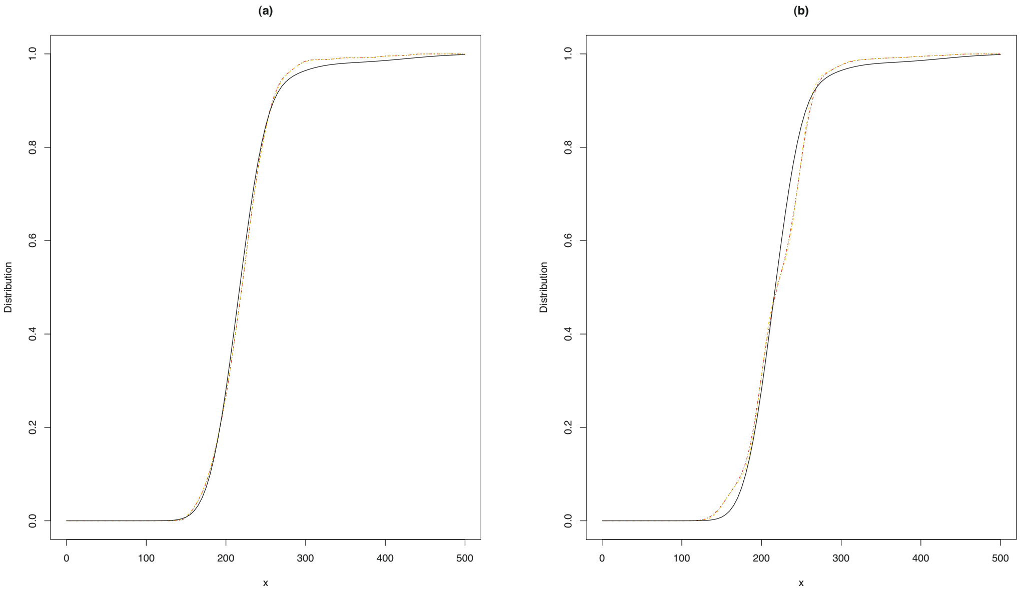

Figure 3: *Kernel estimation using estimator* $\hat{F}_h^g$ *with a sample of size 240. In (a)* $\bar{l} = 5$ *and bandwidths* $h_{MISE_g} = 7.03$ *(red dashed-dotted line),* $\hat{h}_{PB_g} = 8.87$ *(yellow dashed line), and* $h^*_{MISE} = 7.14$ *(green dotted line). In (b),* $\bar{l} = 50$ *and* $h_{MISE_g} = 14.71$ *(red dashed-dotted),* $\hat{h}_{PB_g} = 12.72$ *(yellow dashed line), and* $h^*_{MISE} = 14.82$ *(green dotted line). In both, black solid lines represent the mixture reference distribution.*

In this case, instead of considering a data set already collected in a grouped way, for instance, the real emergence data set of *Avena sterilis* (wild oat) presented in the main paper, we have preferred to apply the new approaches to different grouped data sets constructed from a specific complete real sample. This allows to verify the behavior of the bandwidth selectors as well as the distribution estimator in a concrete situation with a fixed sample size and under different grouping scenarios. To do this, first, we grouped the data in intervals, considering sets of intervals with different average interval lengths. As a grouping measure, the ratio $\omega = \bar{l}/r$, with r denoting the data range, was used. Note that, in this case, we do not know the theoretical distribution function. So, for the sake of comparison, we considered as a benchmark distribution that provided by the standard kernel distribution estimator $\hat{F}_h$ for complete data, given in equation (2) of the main paper. To compute $\hat{F}_h$, the Polansky and Baker bandwidth, $\hat{h}_{PB}$, and the bootstrap bandwidth with complete data, h^*_{boot}, for complete data, were used. Almost identical results were obtained in both cases.

Figure 4 shows the logarithm of the ratio $\text{ISD}\left(\hat{F}^g_{\hat{h}_{PB_g}}\right)/\text{ISD}\left(\hat{F}^g_{h^*_{MISE}}\right)$, with $\text{ISD}\left(\hat{F}^g_{\hat{h}_{PB_g}}\right)$ being the integrated squared distance of $\hat{F}^g_{\hat{h}_{PB_g}}$ with respect to $\hat{F}_{\hat{h}}$,

$$\text{ISD}\left(\hat{F}^g_{\hat{h}_{PB_g}}\right) = \int \left[\hat{F}^g_{\hat{h}_{PB_g}}(u) - \hat{F}_{\hat{h}}(u)\right]^2 du,$$

where $\hat{F}_{\hat{h}}$ is the standard kernel distribution estimator for complete data, using $\hat{h} = \hat{h}_{PB}$ or $\hat{h} = h^*_{boot}$. A similar definition is used for $\text{ISD}\left(\hat{F}^g_{h^*_{MISE}}\right)$, but changing $\hat{h}_{PB_g}$ by h^*_{MISE}. The left panel shows the results obtained using the Polansky and Baker bandwidth $\hat{h}_{PB}$, and the right panel those when employing the kernel distribution estimator with bootstrap bandwidth, h^*_{boot}, as a reference.

Both figures present a very similar pattern, no matter if the standard kernel distribution estimator for complete data was used with $\hat{h} = \hat{h}_{PB}$ or with $\hat{h} = h^*_{boot}$ as a reference. It can be observed that, in general, for all ω values, the logarithms of the ratios are larger than zero or, in other words, $\mathrm{ISD}\left(\hat{F}^g_{\hat{h}_{PB_g}}\right)$ is, in general, larger than $\mathrm{ISD}\left(\hat{F}^g_{h^*_{MISE}}\right)$, showing a better performance of h^*_{MISE} with respect to $\hat{h}_{PB_g}$. Note also that only for some very large values of ω, very high values of the ratio were obtained. Therefore, it was necessary to consider a really heavy grouped data case to start noticing the poor performance of the bandwidth selector $\hat{h}_{PB_g}$ as well as the estimator $\hat{F}^g_h$, with respect to using h^*_{MISE}. This seems to confirm the idea that even though the selected bandwidth is not too close to the optimal, the effect on the distribution estimate is not that severe, unless the grouping effect is really heavy.

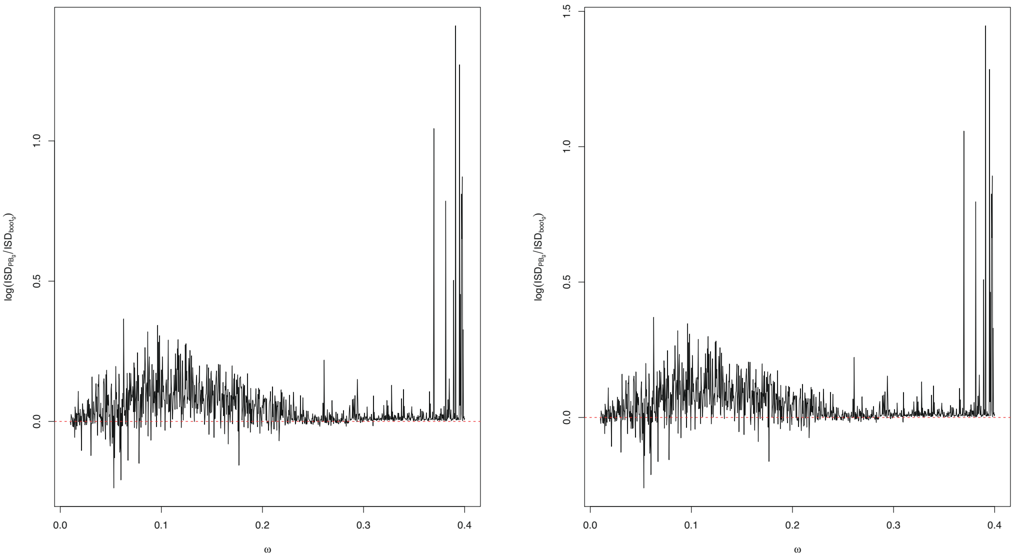

Figure 4: *Natural logarithm of* $\mathrm{ISD}\left(\hat{F}^g_{\hat{h}_{PB_g}}\right)/\mathrm{ISD}\left(\hat{F}^g_{h^*_{MISE}}\right)$ *using* $\hat{F}_{\hat{h}_{PB}}$ *(left panel) and* $\hat{F}_{\hat{h}_{boot_g}}$ *(right panel) as a reference, versus* $\omega = \bar{l}/r$.

To check the effect on the distribution estimator $\hat{F}^g_h$ of the bandwidth selectors $\hat{h}_{PB_g}$ and h^*_{MISE}, the distribution estimates using $\hat{F}_h$ with ungrouped data and $\hat{h}_{PB}$, as well as $\hat{F}^g_h$ with $\hat{h}_{PB_g}$ and h^*_{MISE} for different values of ω are presented in Figure 5. It can be observed that even though the smoothing parameters are quite different, the shape of the distribution estimates are practically identical. This shows from another perspective that distribution estimation is a resistant procedure against grouping effects.

Considering that, in practice, the best results are obtained when the data are not grouped and the standard kernel distribution estimator $\hat{F}_h$ is used, we performed a new simulation experiment to evaluate the average performance of the estimator $\hat{F}^g_h$ (using $\hat{h}_{PB_g}$ and h^*_{MISE}) with respect to $\hat{F}_h$ using $\hat{h}_{PB}$ (similar results were obtained when using h^*_{boot} instead of $\hat{h}_{PB}$). For this, resamples with replacement were taken from the original data and were grouped using different values of ω. Then, $\rho = \mathrm{ISD}\left(\hat{F}^g_h\right)/\mathrm{MISD}\left(\hat{F}_{\hat{h}_{PB}}\right)$ was calculated, for $h = \hat{h}_{PB_g}$ and $h = h^*_{MISE}$, where MISD

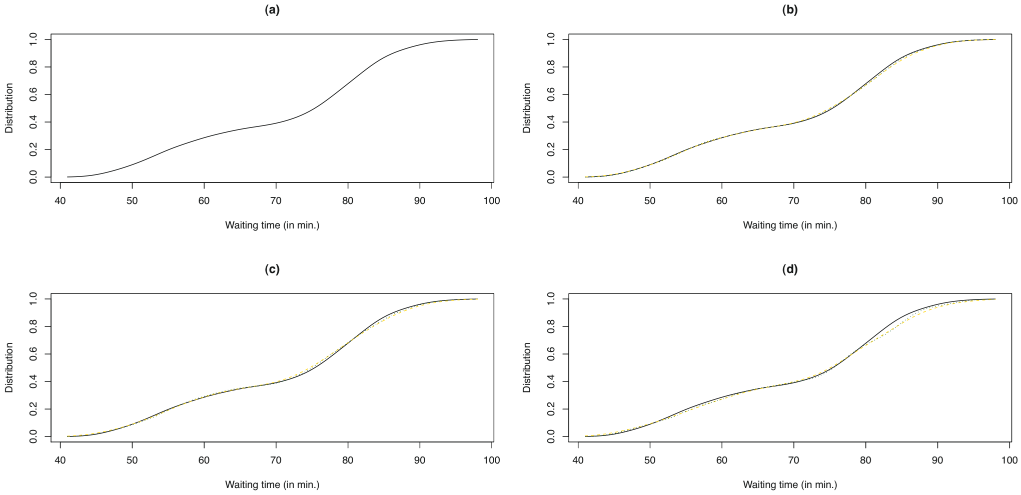

Figure 5: *Kernel distribution estimation: (a) using the standard estimator* $\hat{F}_h$ *with ungrouped data and* $\hat{h}_{PB} = 2.01$*; (b) using* $\hat{F}_h^g$ *with* $\omega = 0.05$, $\hat{h}_{PB_g} = 2.56$ *and* $h^*_{MISE} = 1.72$*; (c) using* $\hat{F}_h^g$ *with* $\omega = 0.08$, $\hat{h}_{PB_g} = 3.01$ *and* $h^*_{MISE} = 1.87$*; (d) using* $\hat{F}_h^g$ *with* $\omega = 0.15$, $\hat{h}_{PB_g} = 3.66$ *and* $h^*_{MISE} = 2.56$*. In all four cases, the solid line represents the kernel distribution estimation using* $\hat{F}_h$ *(ungrouped data), while the yellow dashed lines are the kernel distribution estimations using* $\hat{F}_h^g$ *(grouped data) and* $\hat{h}_{PB_g}$*, and the green dotted lines the kernel distribution estimations using* $\hat{F}_h^g$ *(grouped data) and* h^*_{MISE}*.*

denotes the mean integrated squared distance, given by

$$\mathrm{MISD}\left(\hat{F}^g_{\hat{h}_{PB_g}}\right) = \mathbb{E}\int\left[\hat{F}^g_{\hat{h}_{PB_g}}(u) - \hat{F}_{\hat{h}}(u)\right]^2 du.$$

In order to obtain the average trend, the process was repeated $B = 1000$ times.

Figure 6 shows the result of computing $\log_{10}\rho$ for a fine grid of values of ω. Left panel presents the results for $\hat{h}_{PB_g}$ and right panel those for h^*_{MISE}. What ρ measures is how much the error of the distribution estimate is increasing due to the degree of grouping ω. Both plots show a similar pattern, for small values of ω up to $\omega \approx 0.10$, the estimates perform on average as if there were no grouping. Beyond this value is where the grouping effect becomes important. Yet, note that for $\omega = 0.20$ (i.e., around 5 intervals), $\log_{10}\rho \approx 0.3366$ on average, which means that the error of the estimates merely increased by around $10^{0.3366} \approx 2.17$ times due to the grouping effect, and for very heavy grouping, like $\omega = 0.27$ (i.e., around just 4 intervals), the error increased $10^{0.75} \approx 5.6$ times due to the grouping effect.

3 Pilot bandwidth selection in the bootstrap bandwidth

As pointed out in the main paper, an important step in the bootstrap bandwidth selector proposed in Section 4.2 of the manuscript is that of selecting the pilot bandwidth ζ. In the paper, a method inspired by the idea of smoothing splines, based on selecting the pilot parameter that minimizes the squared distance between the nonparametric cdf estimator and the empirical distribution function, plus a penalty term to avoid obtaining very small bandwidths is proposed. Although this approach seems to be a bit convoluted, we have explored other (simpler) alter-

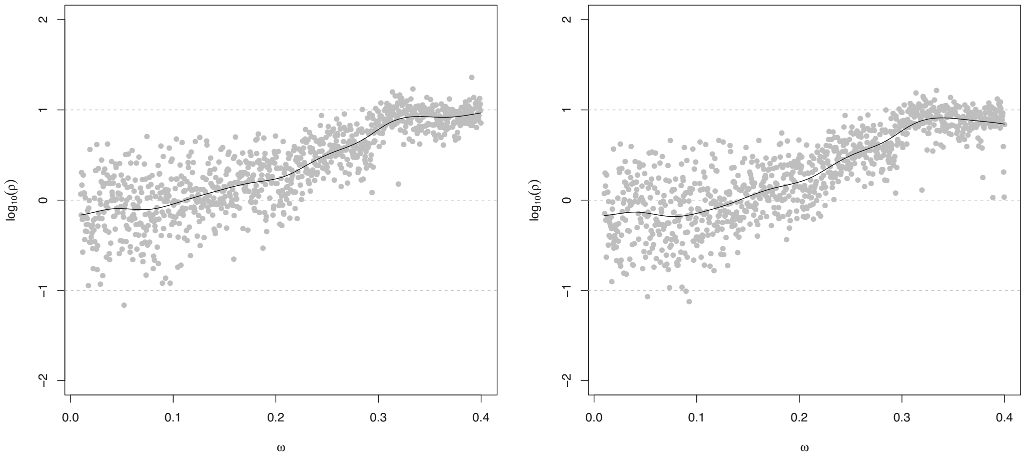

Figure 6: $\log_{10}(\rho)$ *versus* ω *for the Old Faithful geyser data set. Left panel shows the results for* $\hat{h}_{PB_g}$ *and right panel the results for* h^*_{MISE}.

natives to select the pilot bandwidth, obtaining worse results. As an illustration, in this section, some simulation experiments comparing the performance of the bootstrap bandwidth when the pilot bandwidth is selected using the method described in the paper and when it is selected using the plug-in technique are presented.

In this simulation study, we used the function `bw.dist.binned.boot` included in the `R` package `binnednp`. This library was developed by the authors of this paper, jointly with a weed scientist and two computer engineers, and contains some functions implementing most of the nonparametric methods for grouped data (and related problems), studied by the authors in this and in previous papers. Specifically, the function `bw.dist.binned.boot` allows to compute the bootstrap bandwidth following the approach described in the main paper. As an argument of this function, the user can choose the way of selecting the corresponding pilot bandwidth. This can be a fixed value chosen by the user, or a parameter automatically selected either by the method described in the paper or using a plug-in method.

As a first model in this comparison, we used a normal mixture $f(x)=\sum_{i=1}^{3}\alpha_i\phi_{\mu_i,\sigma_i}$, with $\phi_{\mu,\sigma}$ a $N\left(\mu,\sigma^2\right)$ density, $\alpha=(0.33,0.33,0.33)$, $\mu=(14,22,30)$ and $\sigma=(2.5,2.5,2.5)$, where α, μ and σ are the mixture weights, means and standard deviations, respectively. A sample size of $n=200$ was considered. Additionally, we used two grouping degrees, fixing the number of intervals in $k=10$ and $k=30$. In Figure 7, the box-plots of the ratios between the bootstrap bandwidth, h^*_{MISE}, given in equation (17) of the main paper, and the MISE optimal bandwidth, h_{MISE_g}, are shown. Left box-plots correspond to the bootstrap bandwidths when, as pilot bandwidth, we use the plug-in bandwidth, $\hat{h}_{PB_g}$, given in equation (15) of the paper. Right box-plots present the case of using ζ^{λ}_{emp}, obtained as described in Section 4.2 of the main paper, as the pilot bandwidth. Box-plots in left panel show the results for $k=10$ intervals and those in right panel the ones for $k=30$ intervals.

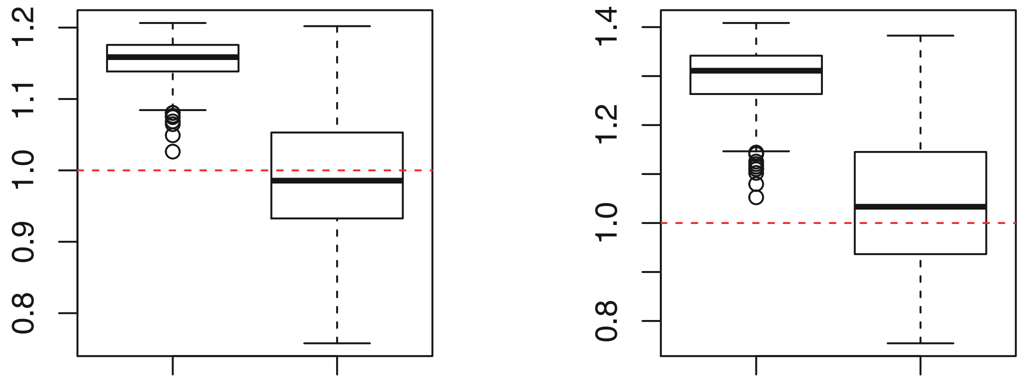

Figure 7: *Box-plots for* $\frac{h^*_{MISE}}{h_{MISE_g}}$. *Left box-plots correspond to the bootstrap bandwidths when, as pilot bandwidth, we use the plug-in bandwidth,* $\hat{h}_{PB_g}$. *Right box-plots present the case of using* ζ^{λ}_{emp} *as the pilot bandwidth. Left panel shows the results for* $k = 10$ *and right panel for* $k = 30$ *intervals.*

In the second part of the study, we used exactly the same model and the same grouping scenarios as those described in detail in the main paper. This means that, as the population density, we used a normal mixture $f(x) = \sum_{i=1}^{4} \alpha_i \phi_{\mu_i,\sigma_i}$, with $\phi_{\mu,\sigma}$ a $N\left(\mu,\sigma^2\right)$ density, $\alpha = (0.70, 0.22, 0.06, 0.02)$, $\mu = (207, 237, 277, 427)$ and $\sigma = (25, 20, 35, 50)$. Moreover, two different scenarios were considered based mainly on Assumption 3.4 of the paper.

S1. $n^{\frac{5}{9}} \bar{l} \to 0$
S2. $n^{\frac{5}{9}} \bar{l} \to \infty$

The set of intervals were also chosen following the same steps as in the main paper, trying to mimic the asymptotic conditions on $\bar{l}$ in Assumption 3.4.

Figures 8 and 9 show the results of the ratios $\frac{h^*_{MISE}}{h_{MISE_g}}$ as box-plots, for scenarios S1 and S2, respectively. In both figures, left box-plots show the results when, as pilot bandwidth, the plug-in bandwidth, $\hat{h}_{PB_g}$, is used, while the right box-plots present the results when, as pilot bandwidth, the selector ζ^{λ}_{emp} is employed.

In all these experiments, a better performance of the bootstrap selector is observed when using the pilot bandwidth obtained with the method inspired by smoothing splines.

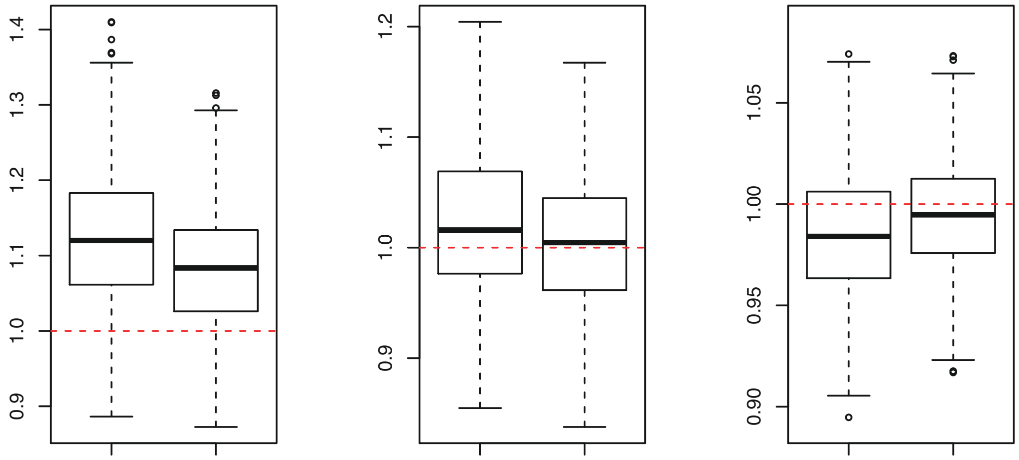

Figure 8: *Box-plots for* $\frac{h^*_{MISE}}{h_{MISE_g}}$ *in scenario S1. Left box-plots correspond to the bootstrap bandwidths when, as pilot bandwidth, we use the plug-in bandwidth,* $\hat{h}_{PB_g}$*. Right box-plots present the case of using* ζ^{λ}_{emp} *as the pilot bandwidth. Left panel shows the results for* $n = 60$*, central panel for* $n = 240$*, and right panel for* $n = 960$*.*

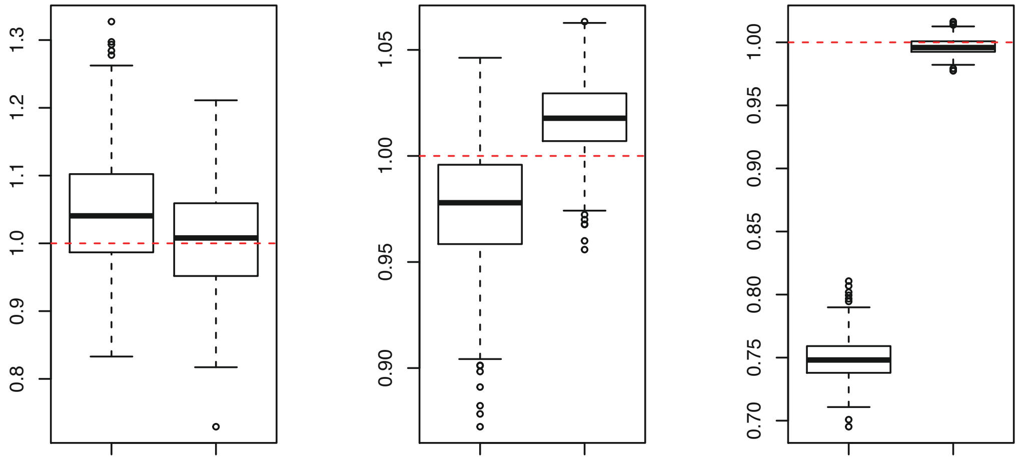

Figure 9: *Box-plots for* $\frac{h^*_{MISE}}{h_{MISE_g}}$ *in scenario S2. Left box-plots correspond to the bootstrap bandwidths when, as pilot bandwidth, we use the plug-in bandwidth,* $\hat{h}_{PB_g}$*. Right box-plots present the case of using* ζ^{λ}_{emp} *as the pilot bandwidth. Left panel shows the results for* $n = 60$*, central panel for* $n = 240$*, and right panel for* $n = 960$*.*