SORT 43 (2) July-December 2019, 259-288 DOI: 10.2436/20.8080.02.88

Kernel distribution estimation for grouped data

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

Abstract

Interval-grouped data appear when the observations are not obtained in continuous time, but monitored in periodical time instants. In this framework, a nonparametric kernel distribution estimator is proposed and studied. The asymptotic bias, variance and mean integrated squared error of the new approach are derived. From the asymptotic mean integrated squared error, a plug-in bandwidth is proposed. Additionally, a bootstrap selector to be used in this context is designed. Through a comprehensive simulation study, the behaviour of the estimator and the bandwidth selectors considering different scenarios of data grouping is shown. The performance of the different approaches is also illustrated with a real grouped emergence data set of *Avena sterilis* (wild oat).

MSC: 62G05, 62N99, 62G09.

Keywords: Bootstrap bandwidth, cumulative distribution function estimator, interval data, plug-in bandwidth

1 Motivation

In the experimental sciences, data usually come from measurements of continuous variables such as temperature, mass, weight, time, length, etc. However, for several reasons, measurements are always obtained in finite precision; i.e., all observed data are rounded or grouped to some extent.

A typical situation in which grouped data clearly appear (and the degree of grouping can be considerable) is when researchers observe variables not continuously, but periodically, thus obtaining time to event data distributed along a set of consecutive intervals. Situations like this appear very frequently in areas such as engineering, economics, social sciences, epidemiology, medicine, agriculture and more (Coit and Dey, 1999, Guo, 2005, Minoiu and Reddy, 2009, Pipper and Ritz, 2007, Rizzi et al., 2016). Especially in these cases, data uncertainty should be taken into account to avoid serious mistakes when making inferences.

[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

Received: February 2019
Accepted: July 2019

One of these situations that partially motivated this work was a real problem from weed science, where weed emergence is coded as a set of non-equally spaced grouped data. In this framework, a key variable to study seedling emergence is the cumulative hydrothermal time (CHTT), which is a mix of time of exposure to certain temperature and humidity conditions. CHTT is typically available in k inspections and the number of emerged seedlings at each one is registered. In more restrictive situations, only the cumulative proportion of emerged seedlings recorded at every monitoring date are reported. Most of the statistical methods used in this context tackle the problem of modeling weed emergence (the so-called emergence curve) from a regression point of view. Parametric models such as Gompertz and logistic have been widely used to define the relationship between the CHTT and weed emergence. However, due to the limitations of this approach, in Cao et al. (2013), this problem has been dealt with through nonparametric estimation of the distribution function of the CHTT at emergence. In that paper, a simple kernel distribution estimator adapted to deal with grouped data, based on a modification of the standard kernel estimator of the distribution function, was proposed and applied to analyse a weed emergence data set. This nonparametric approach has recently been proven to outperform the classical regression methods in terms of prediction error (González-Andújar et al., 2016). However, a deeper statistical analysis of this new nonparametric distribution estimator is required. In the present paper, we study the asymptotic properties of this estimator. Additionally, a plug-in and a bootstrap bandwidth selector are proposed and compared in different scenarios through a comprehensive simulation study.

The organization of this paper is as follows. In Section 2 the notation used throughout the paper and the kernel distribution estimator for grouped data are presented. In Section 3, under some assumptions, the asymptotic bias, variance, and mean integrated squared error (MISE) of this estimator are obtained. In Section 4, using the asymptotic MISE expression, a plug-in bandwidth selector is proposed. Additionally, closed forms for the MISE and its bootstrap version, MISE*, are presented and a bootstrap bandwidth selector is derived. In Section 5, a simulation study with different sample sizes is presented to show the consistency of the estimator under different grouping scenarios. In Section 6, the nonparametric estimator and both bandwidth selection methods are applied to a grouped emergence data of *Avena sterilis* (wild oat). Finally, Section 7 summarizes the main conclusions. Proofs are included in Appendix A. Supplementary materials completing the simulation study and with an additional empirical study based on real data are available online.

2 Kernel distribution estimator for interval-grouped data

Let us introduce the notation for grouped data. Suppose that X is the random variable of interest, with density function f and distribution function F, and let $(X_1, X_2, \dots, X_n)$ be a random sample of X. Consider a set of intervals $[y_{j-1}, y_j)$, $j = 1, 2, \dots, k$, where the j-th interval length is $l_j = y_j - y_{j-1}$, its midpoint is $t_j = \frac{1}{2}(y_{j-1} + y_j)$, and denote the number of observations within each interval by $(n_1, n_2, \ldots, n_k)$. Sometimes, only the sample proportions $(w_1, w_2, \ldots, w_k)$ are available, where $w_j = F_n(y_j-) - F_n(y_{j-1}-)$ is the actual observed random quantity, and $F_n(y-)$ is the left-hand limit of the empirical distribution function F_n.

For example, using this notation and focusing on the weed emergence problem that motivated this research, X would be the random variable measuring the CHTT at emergence of a particular weed. Moreover, denoting by n the number of seedlings that have emerged at the end of the monitoring process, since the inspections carried out to count the number of emerged seedlings are performed at a limited number of k instants, the values $X_1, X_2, \ldots, X_n$, measuring the CHTT at emergence of every single seedling, cannot be observed. In this case, what is observed is the CHTTs at inspections (the limits of the intervals, previously denoted by y_i, $i = 0, 1, \ldots, k$) and the total number of seedlings that have emerged in the intervals between consecutive inspection times, $n_1, n_2, \ldots, n_k$, (or the corresponding sample proportions, $w_1, w_2, \ldots, w_k$, with $w_i = n_i/n$).

It is worth mentioning that there exists a parallelism between grouped data and the so called interval-censored data (see the book by Klein and Moeschberger, 1997, for an introduction about interval-censored data in survival analysis). The main similarity is that the exact value of the interest random variable data X_i is not observed and one is only able to know the interval in which every datum of the interest population belongs. There are two main differences between grouped data and interval-censored data. The first one is that the intervals $[y_{j-1}, y_j)$ are typically fixed (not random) for grouped data, while the interval endpoints are random variables for interval-censored data. As a consequence, for interval-censored data there are, in principle, as many different intervals as the sample size, n, while for grouped data the number of different intervals, k, is known beforehand and is smaller than the sample size ($k < n$). General estimation methods applicable for interval-censored data, as Turnbull's estimator (Turnbull, 1976), can also be used for grouped data. In our grouped data setup, Turnbull's estimator of the cumulative distribution function just gives the empirical cumulative distribution function for grouped data.

First, let us consider the ideal continuous case, where $(X_1, X_2, \ldots, X_n)$ are supposed to be observed. From the well-known Parzen-Rosenblatt kernel density estimator (Parzen, 1962, Rosenblatt, 1956), defined as

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - x_i), \tag{1}$$

where $K_h(u) = h^{-1} K(u/h)$, with K a kernel function (typically an auxiliary density function) and h the bandwidth parameter, it is straightforward to obtain a kernel estimator of the cumulative distribution function (cdf) as

$$\hat{F}_h(x) = \int_{-\infty}^{x} \hat{f}_h(t)\,dt = \frac{1}{n}\sum_{i=1}^{n}\mathbb{K}\left(\frac{x-x_i}{h}\right), \tag{2}$$

where $\mathbb{K}(t) = \int_{-\infty}^{x} K(t)\,dt$. Some theoretical properties of (2) can be found in Hill (1985), Nadaraya (1964) and Reiss (1981). Although the choice of the kernel function is of secondary importance, the bandwidth h plays a crucial role in the shape of estimator (2). Differently to the case of kernel density estimation, there are not many contributions addressing the bandwidth selection problem in kernel distribution estimation. Different cross-validation methods were studied in Bowman, Hall, and Prvan (1998) and Sarda (1993); and plug-in selectors were considered in Altman and Leger (1995) and Polanski and Baker (2000). The interested reader can find more theoretical details and an extended discussion on the previous cross-validation and plug-in selectors in Quintela-del-Río and Estévez-Pérez (2012). More recently, a bootstrap bandwidth selector for the estimator (2) has been developed in Dutta (2015).

When working with grouped data, the issue of density estimation has been widely addressed employing different approaches. Using nonparametric methods, in Reyes, Francisco-Fernandez, and Cao (2016) a simple modification of the estimator (1) was proposed and studied, while in Reyes, Francisco-Fernández, and Cao (2017) two different bandwidth selectors for this estimator were analysed. Also in a nonparametric context, Wang and Wertelecki (2013) proposed a bootstrap type kernel density estimator for binned data, and Blower and Kelsall (2002) proposed a nonlinear binned kernel estimator. In this setting, Rizzi et al. (2016) compared the performance of different nonparametric density estimators for grouped data via a simulation study and also using some empirical cancer data. Other approaches in this framework consist in converting the density estimation problem to a regression problem using the root-unroot algorithm (Brown et al., 2010) or using parametric methods (Wang and Wang, 2016). Parametric methods can be useful for heavy grouping if the assumed model is correct, but if this is not true, the results obtained can be wrong.

Studies on estimation methods for the distribution function for grouped data are much scarcer and they are mainly based on the empirical distribution function (Turnbull, 1976). Although the distribution function is closely connected with the density function, in some situations, the data are collected in an accumulated way, making the distribution function the element of interest. This is the case, for example, in the weed emergence problem previously described. Therefore, it is of special concern to develop and study specific distribution function estimators for grouped observations.

Starting with the related density estimation problem, in Scott and Sheather (1985) and Titterington (1983), the kernel density estimator (1) was redefined to be used with binned data, assuming a constant binwidth, as

$$\tilde{f}_h(x) = n^{-1}\sum_{i=1}^{k} n_i K_h(x-t_i). \tag{3}$$

In Reyes et al. (2016), a modified version of (3) considering the general case of different interval lengths, given by

$$\hat{f}_h^g(x) = \frac{1}{h}\sum_{i=1}^{k} w_i K\left(\frac{x-t_i}{h}\right), \tag{4}$$

was studied. Its asymptotic properties were obtained, and a plug-in bandwidth selector was proposed and analysed.

From (4), it is straightforward to obtain a kernel distribution estimator for binned or grouped data as

$$\hat{F}_h^g(x) = \int_{-\infty}^{x} \hat{f}_h^g(u)\,du = \sum_{i=1}^{k} w_i \mathbb{K}\left(\frac{x-t_i}{h}\right), \tag{5}$$

where $\mathbb{K}(x) = \int_{-\infty}^{x} K(z)\,dz$. Note that (5) is a simple modification of (2) for the context of interval-grouped data.

3 Theoretical results

In this section, a closed form for the MISE of the kernel distribution estimator for grouped data (5) is obtained, and its asymptotic properties are derived. Using standard calculations and assuming that $F(y_k) = 1$ and $F(y_0) = 0$, it is easy to prove that the expectation and the variance of $\hat{F}_h^g(x)$ are, respectively,

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) p_i \tag{6}$$

and

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right) p_i(1-p_i) - \frac{2}{n}\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right) p_i p_j, \tag{7}$$

where $p_i = F(y_i) - F(y_{i-1})$.

From (6) and (7), it is straightforward to obtain a closed expression for the MISE of the estimator defined in (5):

$$\text{MISE}\left(\hat{F}_h^g\right) = \mathbb{E}\left\{\int \left[\hat{F}_h^g(x) - F(x)\right]^2 dx\right\} = B + V, \tag{8}$$

where,

$$B = \int \left\{\mathbb{E}\left[\hat{F}_h^g(x)\right] - F(x)\right\}^2 dx \tag{9}$$

denotes the integrated squared bias and

$$V = \int \mathbb{V}\left[\hat{F}_h^g(x)\right] dx \tag{10}$$

is the integrated variance.

The asymptotic bias and variance of (5) are stated in Theorem 3.1, whose proof is included in Appendix A. The following assumptions are needed.

Assumption 3.1 *The kernel K is a symmetric probability density function with support in $[-1,1]$, at least 3-times differentiable and such that $K^{(3)}$ is bounded.*

Assumption 3.2 *The distribution F has compact support $[\mathscr{L},\mathscr{U}]$, it is 4-times differentiable and $F^{(4)}$ is continuous.*

Assumption 3.3 *The bandwidth $h = h_n$ is a non random sequence of positive numbers such that $\lim_{n\to\infty} h = 0$ and $\lim_{n\to\infty} nh = \infty$.*

Assumption 3.4 *Given a set of $k = k_n$ intervals $[y_{j-1}, y_j)$, $j = 1,2,\ldots,k$, $y_0 \leqslant \mathscr{L}$ and $y_k \geqslant \mathscr{U}$, the average interval length is $\bar{l} = \bar{l}_n = \frac{1}{k}\sum_{i=1}^{k} l_i$, where l_i is the abbreviated notation for the i-th interval length $l_{i,n}$. It is assumed that $\lim_{n\to\infty} \bar{l} = 0$, $\lim_{n\to\infty} n\bar{l} = \infty$, $\bar{l} = o\left(h^{5/3}\right)$, and $\max_i \left|l_i - \bar{l}\right| = \max_{1\leqslant i\leqslant k} \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.*

Assumptions 3.1 and 3.2 are just smoothness and differentiability conditions about the kernel K and the distribution function F. Assumption 3.3 is the typical one used in kernel estimation concerning the sample size n and the bandwidth h. However, Assumption 3.4 is of special importance and deserves some comments.

Condition $\lim_{n\to\infty} \bar{l} = 0$ simply states that, as the sample size increases, the average interval length shrinks. This means that, taking into account the condition $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$, all intervals are shrinking as well. However, $\lim_{n\to\infty} n\bar{l} = \infty$ states that n should increase faster than $\bar{l}$ decreases. This is an important condition from a theoretical point of view, as if the intervals shrink faster than n increases, at some point there would be more intervals than data points, and some of the intervals would be empty or there would not be enough data points in each interval.

Condition $\bar{l} = o\left(h^{5/3}\right)$ states an intuitive idea: as the sample size n increases, the average length $\bar{l}$ must vanish faster than, at least, h (concretely, faster than $h^{5/3}$). This condition has a practical basis. Since the average distance between points is $\bar{l}$, the bandwidth must be greater than $\bar{l}$ at all times to gather information from the surroundings. In other words, as n increases, h must vanish, but always behind $\bar{l}$.

Regarding the condition about $\max_i \left|l_i - \bar{l}\right|$, at first, this is necessary from the strictly mathematical viewpoint, but in practice it is a way for controlling the variability of the intervals. This assumption means that the lengths of the intervals are not very different. In other words, in our assumptions we unquestionably accept different interval lengths in order to generalize the binned estimator, but within certain limits, and these limits of maximum variability are controlled by $\bar{l}$ via $\max_i \left|l_i - \bar{l}\right| = o\left(\bar{l}\right)$.

Theorem 3.1 *Under Assumptions 3.1 to 3.4,*

$$\mathrm{MSE}\left[\hat{F}_h^g(x)\right] = \frac{h^4}{4}\mu_2(K)^2 F''(x)^2 + \frac{1}{n}F(x)\left[1-F(x)\right] - \frac{h}{n}F'(x)C_0 + o\left(\frac{h}{n}\right) + o\left(h^4\right)$$

and

$$\mathrm{MISE}\left(\hat{F}_h^g\right) = \mathrm{AMISE}\left(\hat{F}_h^g\right) + o\left(\frac{h}{n}\right) + o\left(h^4\right),$$

where

$$\mathrm{AMISE}\left(\hat{F}_h^g\right) = \frac{h^4}{4}\mu_2(K)^2 A\left(f'\right) + \frac{1}{n}\int F(x)\left[1-F(x)\right]dx - \frac{h}{n}C_0 \tag{11}$$

with $A\left(f'\right) = \int f'(x)^2 dx$, and

$$C_0 = 2\int zK(z)\mathbb{K}(z)\,dz > 0.$$

Remark 3.1 *Since the distribution function F has compact support (Assumption 3.2) then the integral $\int F(x)\left(1-F(x)\right)dx$ is finite. This needs not be the case for a general cdf F.*

Remark 3.2 *Taking care of higher order terms in the asymptotic expansions for the MSE and the MISE, the resulting approximations show the impact of the average interval length, $\bar{l}$, in these error criteria:*

$$\begin{aligned}
\mathrm{MSE}\left[\hat{F}_h^g(x)\right] &= \left(\frac{h^2}{2}\mu_2(K) + \frac{\bar{l}^2}{12}\right)^2 F''(x)^2 + \frac{1}{n}F(x)\left[1-F(x)\right] - \frac{h}{n}F'(x)C_0 \\
&+ \frac{\bar{l}^2}{24n}F''(x) + o\left(\frac{h}{n}\right) + o\left(h^4\right) + o\left(\bar{l}^4\right) + o\left(h^2\bar{l}^2\right) + o\left(\frac{\bar{l}^2}{n}\right) \\
\mathrm{MISE}\left(\hat{F}_h^g\right) &= \left(\frac{h^2}{2}\mu_2(K) + \frac{\bar{l}^2}{12}\right)^2 A\left(f'\right) + \frac{1}{n}\int F(x)\left[1-F(x)\right]dx - \frac{h}{n}C_0 \\
&+ o\left(\frac{h}{n}\right) + o\left(h^4\right) + o\left(\bar{l}^4\right) + o\left(h^2\bar{l}^2\right) + o\left(\frac{\bar{l}^2}{n}\right).
\end{aligned}$$

Under Assumptions 3.1 - 3.4, these two expressions reduce to the asymptotic expressions given in Theorem 3.1.

4 Bandwidth selectors

As pointed out in Section 1, the kernel distribution estimator (2) heavily depends on the bandwidth h. Obviously, the same occurs for the estimator adapted for grouped data (5), since too small bandwidths give estimates that are too close to the empirical cdf, and too large selections tend to provide oversmoothed estimators. In this sense, it is very important to have an automatic bandwidth selection method producing reliable estimates for a real data set. In this section, two bandwidth selectors (plug-in and bootstrap) are proposed for (5) in the context of interval-grouped data.

4.1 Plug-in bandwidth selector

From Eq. (11), it is immediate to get an asymptotically optimal global bandwidth. Taking the first derivative of (11), equating to zero and solving for h, it follows that

$$h_{AMISE} = \left[\frac{C_0}{n\mu_2 (K)^2 A(f')} \right]^{\frac{1}{3}}. \tag{12}$$

Note that Eq. (12) is the same as that for continuous data (see, e.g., Azzalini, 1981, Hill, 1985, Mack, 1984). However, it is important to keep in mind that (12) holds as an asymptotic optimal bandwidth for grouped data only as long as Assumptions 3.1 to 3.4 hold. Otherwise, some other important terms of the asymptotic expansion of MISE $\left(\hat{F}_h^g\right)$ remain non-negligible, thus making (11) fall short as a MISE $\left(\hat{F}_h^g\right)$ approximation.

In Eq. (12), an estimate of $A(f')$ is required to have a practical bandwidth. To estimate $A(f')$, we used the proposal of Polansky and Baker (2000) adapted for grouped data. Other approaches could be used here, but we preferred the Polansky and Baker method for computational reasons and because it gave stable results when using grouped data. In the continuous data case, Polansky and Baker (2000) proposed to estimate $A(f')$ by $-\hat{\psi}_{\eta,2}$, where

$$\hat{\psi}_{\eta,2} = \frac{1}{n^2\eta^3} \sum_{i=1}^{n} \sum_{j=1}^{n} L'' \left(\frac{X_i - X_j}{\eta} \right), \tag{13}$$

L being a kernel function (possibly different from K) and $\eta > 0$ an auxiliary smoothing parameter. The bandwidth η can be selected using a plug-in procedure. For this, it would be necessary to obtain the asymptotic MSE of $\hat{\psi}_{\eta,2}$, that depends on $\psi_4 = \int f^{(4)}(x) f(x) dx$, and then estimate ψ_4. Clearly, the problem still remains, since estimating ψ_4 will depend on an initial bandwidth, which in turn will depend on $\psi_6 = \int f^{(6)}(x) f(x) dx$, and so on. A common strategy is to estimate ψ_u with some quick and simple rule, like the normal scale rule (Wand and Jones, 1995). Once $\hat{\psi}_{\eta,u}$ is obtained, it is possible to select a bandwidth for estimating ψ_{u-2}. Then, having estimated $\hat{\psi}_{\eta,u-2}$, a bandwidth for estimating ψ_{u-4} can be selected, and so forth. Polansky and Baker (2000) suggest using the same iterative method.

In the context of grouped data, we propose to estimate $A(f')$ with $\hat{A}_{PB_g} = -\hat{\psi}^g_{\eta,2}$, where $\hat{\psi}^g_{\eta,2}$ is an appropriate version of (13), given by:

$$\hat{\psi}^g_{\eta,2} = \frac{1}{\eta^3} \sum_{i=1}^{k} \sum_{j=1}^{k} L'' \left(\frac{t_i - t_j}{\eta} \right) w_i w_j. \tag{14}$$

Similar steps to those described previously for continuous data can be followed now to select the bandwidth η. It should be noted that in this case, to obtain a plug-in bandwidth for η it is necessary to derive the asymptotic MSE of $\hat{\psi}^g_{\eta,2}$ using grouped data. In Reyes et al. (2017), both the asymptotic variance and bias of $\hat{\psi}^g_{\eta,u}$ were derived for $u > 0$. Based on those, a way of selecting the plug-in bandwidth for $\hat{\psi}^g_{\eta,u}$ was proposed. Using that approximation with $u = 2$ and plugging $\hat{A}_{PB_g}$ into (12) gives a practical plug-in bandwidth selector for $\hat{F}^g_h(x)$,

$$\hat{h}_{PB_g} = \left[\frac{C_0}{n\mu_2(K)^2 \hat{A}_{PB_g}} \right]^{\frac{1}{3}}. \tag{15}$$

Note that using similar arguments to those employed in Theorem 2 of Reyes et al. (2017), the relative rate of convergence for the plug-in bandwidth $\hat{h}_{PB_g}$ can be derived.

4.2 Bootstrap bandwidth selector

The bootstrap method can be used to produce an estimator of the MISE. In the grouped data setup, this has been already proposed by Reyes et al. (2017) for density estimation. These authors have proved that there exists a closed expression for the bootstrap version of the MISE in that context. This implies that Monte Carlo is not needed to obtain a bootstrap approximation of the MISE in density estimation for grouped data. This will be also the case for cdf estimation for grouped data.

To build a bootstrap version of the MISE, we consider a pilot bandwidth, ζ, and construct the grouped-data smooth estimator of F as defined in (5), but replacing h by ζ. The idea is to draw resamples from $\hat{F}^g_\zeta$, to group the data and to compute the estimator $\hat{F}^g_h$ with those bootstrap samples. The bootstrap resampling plan proceeds as follows.

1. Fix some pilot bandwidth, ζ, and consider the grouped-data smooth cdf estimator, $\hat{F}^g_\zeta$.
2. Draw $(n_1^*, \ldots, n_k^*)$ from a multinomial distribution $\mathscr{M}_k(n; \tilde{p}_1^\zeta, \ldots, \tilde{p}_k^\zeta)$, with $\tilde{p}_i^\zeta = \hat{F}^g_\zeta(y_i) - \hat{F}^g_\zeta(y_{i-1})$, $i = 1, \ldots, k$, and define $w_i^* = n_i^*/n$.

3. Compute the grouped-data smooth cdf estimator based on this bootstrap resample:

$$\hat{F}_h^{g*}(x) = \sum_{i=1}^{k} w_i^* \mathbb{K}\left(\frac{x-t_i}{h}\right).$$

4. Define the bootstrap version of MISE:

$$\text{MISE}^*\left(\hat{F}_h^{g*}\right) = \mathbb{E}^*\left\{\int \left[\hat{F}_h^{g*}(x) - \hat{F}_\zeta^g(x)\right]^2 dx\right\},$$

   where, $\mathbb{E}^*$ denotes the bootstrap expectation (with respect to $\hat{F}_\zeta^g$).

Remark 4.1 *Since, under Assumption 3.1, the support of $\hat{F}_\zeta^g$ is $[t_1 - \zeta, t_k + \zeta]$, it may happen that this interval is not contained in $[y_0, y_k]$. This only happens if $\zeta \leq \frac{1}{2}\min\{l_1, l_k\}$. So, in order to resample from a distribution with support contained in $[y_0, y_k]$, we consider the conditional distribution corresponding to $\hat{F}_\zeta^g$ restricted to the interval $[y_0, y_k]$.*

The previous remark implies that it may happen that $\sum_{i=1}^k \tilde{p}_i^\zeta < 1$. If this is the case, we define

$$\hat{p}_i^\zeta = \frac{\tilde{p}_i^\zeta}{\sum_{j=1}^k \tilde{p}_j^\zeta},\ i = 1, 2, \ldots, k, \tag{16}$$

and we draw the bootstrap resamples in Step 2 from a multinomial distribution with probabilities $\hat{p}_i^\zeta$.

Substituting p_i by $\hat{p}_i^\zeta$ in (6) and (7), the bootstrap version of the mean integrated squared error admits the following closed expression:

$$\text{MISE}^*\left(\hat{F}_h^{g*}\right) = \mathbb{E}^*\left\{\int \left[\hat{F}_h^{g*}(x) - \hat{F}_\zeta^g(x)\right]^2 dx\right\} = B^* + V^*,$$

where

$$B^* = \int \left\{\mathbb{E}^*\left[\hat{F}_h^{g*}(x)\right] - \hat{F}_\zeta^g(x)\right\}^2 dx$$

and

$$V^* = \int \mathbb{V}^*\left[\hat{F}_h^{g*}(x)\right] dx,$$

with

$$\mathbb{E}^*\left[\hat{F}_h^{g*}(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right)\hat{p}_i^\zeta$$

and

$$\mathbb{V}^*\left[\hat{F}_h^{g*}(x)\right] = \frac{1}{n}\sum_{i=1}^{k} \mathbb{K}^2\left(\frac{x-t_i}{h}\right)\hat{p}_i^\zeta\left(1-\hat{p}_i^\zeta\right) - \frac{2}{n}\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)\hat{p}_i^\zeta\hat{p}_j^\zeta.$$

This approach is computationally efficient since there is no need to use Monte Carlo to approximate the bootstrap resampling distribution. Finally, the bootstrap bandwidth is defined as the minimizer of MISE$^*\left(\hat{F}_h^{g*}\right)$, in the smoothing parameter, h:

$$h^*_{MISE} = \arg\min_{h>0} MISE^*\left(\hat{F}_h^{g*}\right). \tag{17}$$

An important step in this bootstrap procedure is that of selecting the pilot bandwidth ζ. After performing some empirical experiments, we have used 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. The idea consists in finding the parameter, denoted by ζ^λ_{emp}, such that,

$$\zeta^\lambda_{emp} = \arg\min_{h>0} \sum_{i=0}^{k} \left[F_n(y_i) - \hat{F}_h^g(y_i)\right]^2 + \lambda \int \hat{f}_h^{g\prime}(x)^2\,dx,$$

where $\lambda \geq 0$ determines the penalty degree over the global slope of the nonparametric density estimator, defined in (4). To select an "optimal" penalty degree, λ_{opt}, we have used the rule of finding the penalty allowing to obtain a pilot bandwidth that best approximates the overall slope of the population density, that is,

$$\lambda_{opt} = \arg\min_{\lambda\geq 0} \left| A\left(\hat{f}^{g\prime}_{\zeta^\lambda_{emp}}\right) - A(f')\right|.$$

In practice, λ_{opt} can be estimated by

$$\hat{\lambda}_{opt} = \arg\min_{\lambda\geq 0} \left| A\left(\hat{f}^{g\prime}_{\zeta^\lambda_{emp}}\right) - A(\hat{f}'_\theta)\right|,$$

where $\hat{f}'_\theta$ represents a parametric estimator of the first derivative of the density function, fitted with the grouped data sample and flexible enough to capture, at least partially, the global slope of f. It was checked that fitting normal mixture models with a maximum number of $r = 5$ components provided, in general, very good results. In practice, the Expectation Maximization (EM) method (Mclachlan and Peel, 2000) was used to estimate the parameters of these models, using the BIC criterion to select the best fit.

Other simpler alternatives to select the pilot bandwidth, ζ, were also explored, producing in general worse results than those obtained with the algorithm previously described. For that reason (and for reason of space), only the results obtained when using the previous approach to select the pilot bandwidth are shown in the paper. In the Supplementary Materials, some simulations experiments comparing the performance of the bootstrap bandwidth (17) when using as a pilot bandwidth ζ^λ_{emp} and when this auxiliary parameter is derived using the plug-in technique, $\hat{h}_{PB_g}$, are presented. A better performance of the bootstrap selector is clearly observed when using the pilot bandwidth obtained by the method described above.

5 Simulations

To have an idea of the effectiveness of the estimator (5) when using (15) and (17) as bandwidth selectors, some simulation studies were performed under different grouping scenarios. For this, the free statistical software R and packages norlmix and binnednp were used (Barreiro et al., 2019, Mächler, 2017, R Core Team, 2019).

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 $\sigma = (25, 20, 35, 50)$, where α, μ and σ are the mixture weights, means and standard deviations, respectively. This normal mixture was used in weed science to model the relationship between weed emergence of *Bromus diandrus* and hydrothermal time (Cao et al., 2011). A total number of 1000 trials were considered throughout all simulations.

Trying to mimic the asymptotic conditions on $\bar{l}$ in Assumption 3.4, in a first simulation experiment, the behaviour of the MISE for grouped data, denoted by MISE_g, was studied depending on the bandwidth h, for sample sizes 60, 240, and 960. Two different scenarios were considered based mainly on Assumption 3.4.

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

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

In Scenario S1, condition $\bar{l} = o\left(h^{5/3}\right)$ is confirmed for $h \sim n^{-\frac{1}{3}}$ (classical optimal rate in the case of distribution estimation without grouping), for example, h_{AMISE} or $\hat{h}_{PB_g}$; while in Scenario S2 occurs the opposite. It is important to note that in both scenarios $\bar{l}$ tends to zero as n increases.

Note that MISE_g can be approximated by numerical integration in an interval $[a,b]$ using (8), (9) and (10), jointly with the expressions of the expectation and the variance of $\hat{F}_h^g$ in (6) and (7). In practice, we considered $a = 0$, $b = 509.25$. With these values for a and b, the area under the reference normal mixture in $[a,b]$ is 0.999.

To simulate the set of intervals as n increases, the next steps were followed:

1. Consider $\bar{l} = En^{-\alpha}$ and $a_n = Fn^{-\beta}$, where E, α, F and β are positive constants.
2. Take initial interval lengths $\{l_i\}$ for $i = 1, 2, \ldots, 5$: $l_1 = \bar{l} - 4a_n$, $l_2 = \bar{l} + 0.5a_n$, $l_3 = \bar{l} - 1.5a_n$, $l_4 = \bar{l} + 3a_n$, $l_5 = \bar{l} + 2a_n$.
3. For $i > 5$, $l_i = l_{[(i-1) \bmod 5]+1}$, where $[m \bmod \ell]$ stands for m modulo ℓ. Then, the initial set of intervals is repeated one after another, as many times as necessary.

Constants E and F are just selected considering the distribution support. To choose the positive constants α and β, note that according to the initial set of intervals in Step 2, it follows that $\max_i \left|l_i - \bar{l}\right| = 4a_n = 4Fn^{-\beta}$.

Assumption 3.4 and Step 1 impose that $4Fn^{-\beta} = o\left(\bar{l}\right) = o\left(En^{-\alpha}\right)$, which basically is

$$n^{-\beta} = o\left(n^{-\alpha}\right). \tag{18}$$

So, for (18) to hold, $n^{\alpha-\beta} \to 0$, which only occurs when $\alpha - \beta < 0$, i.e., when $\beta > \alpha$.

Now, recall that in Scenario S1, $\bar{l} = o\left(h^{5/3}\right) = o\left(n^{-5/9}\right)$ must hold. Thus, according to Step 1, $\bar{l} = En^{-\alpha} = o\left(n^{-\frac{5}{9}}\right)$, which basically is

$$n^{-\alpha} = o\left(n^{-\frac{5}{9}}\right). \tag{19}$$

This only occurs when $\frac{5}{9} - \alpha < 0$; i.e., when $\alpha > 5/9$.

In brief, to simulate Scenario S1, (18) and (19) must hold, i.e., $\beta > \alpha > 5/9$ must be true. On the other hand, to simulate S2, (18) must hold but (19) must not. It is required that $n^{\frac{5}{9}}\bar{l} \to \infty$, so both $\beta > \alpha$ and $\alpha < 5/9$ must be true. Specifically, in our simulations, we chose $(E,\alpha,F,\beta) = (800,4/5,150,1)$ for S1, and $(E,\alpha,F,\beta) = (37.1,1/20,150,1)$ for S2.

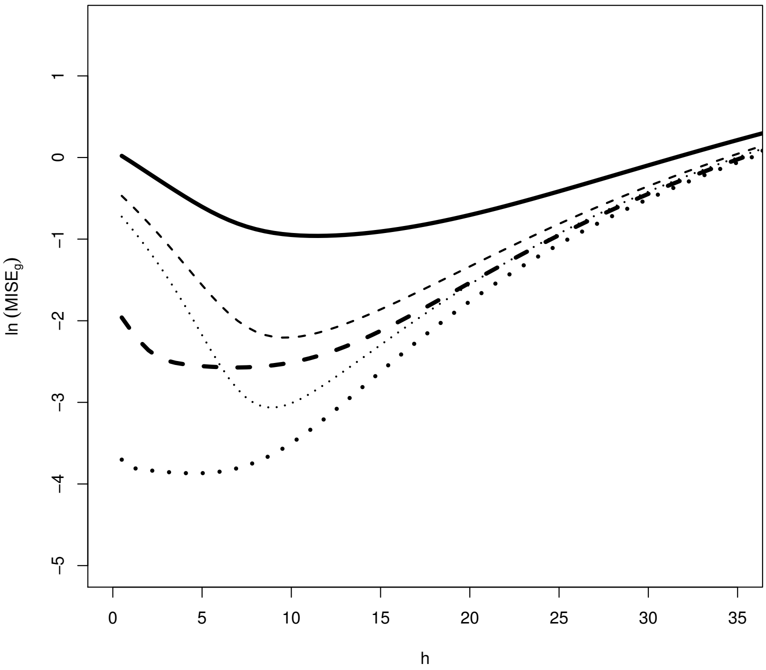

Figure 1: $\ln\left(\text{MISE}_g\right)$ *curves by scenario and sample size. Solid lines are for $n = 60$, dashed lines for $n = 240$ and dotted lines for $n = 960$. Thick lines represent curves in S1, while thin lines represent curves in S2 (note that curves for $n = 60$ are practically the same in both scenarios).*

Firstly, for each sample size and each scenario, MISE$_g$ was approximated over a grid of values of h. Figure 1 shows the MISE$_g$ curves, as a function of h, for the three different sample sizes in both scenarios. Note that a logarithmic scale was used in the vertical axis in order to better appreciate minimal values. This is because very small differences were found in the MISE_g curves for small values of h, particularly for the largest sample size. This suggests that even in the case of grouped data, small deviations from the optimal bandwidth may still give quite good distribution estimates, making the distribution estimation relatively resistant to grouping effects. On the other hand, it is important to note in Figure 1 that MISE_g decreases as the sample size increases, which seems to confirm consistency of the estimator defined in (5). It is also clear that optimal bandwidths for S2 are larger than for S1.

Next, a second simulation experiment was performed to analyse the behaviour of the plug-in bandwidth (15) and the bootstrap selector (17). We compared the sampling distribution of $\hat{h}_{PB_g}$ and h^*_{MISE} with respect to the target values of the bandwidths minimizing MISE_g, denoted by h_{MISE_g}, for each sample size and scenario. The process performed was the following:

1. Simulate an n-size sample from the normal mixture reference density f.
2. Divide the data range into intervals $[y_{i-1}, y_i)$ of length l_i (according to the previous guidelines).
3. Considering the interval midpoints, estimate $A(f')$ by means of $\hat{A}_{PB_g}$ and calculate $\hat{h}_{PB_g}$ using (15).
4. Select ζ as described in Section 4.2 and approximate h^*_{MISE}.
5. Compute $\hat{h}_{PB_g}/h_{MISE_g}$ and h^*_{MISE}/h_{MISE_g}.
6. Repeat Steps 1 to 5 $B = 1000$ times.

Figure 2 shows the results as box-plots. Regarding $\hat{h}_{PB_g}$ (yellow left box-plots for each sample size), it can be observed that starting from the same grouping conditions and sample size, the sampling distribution gets more precise as the sample size increases under both scenarios, S1 and S2. However, in both situations $\hat{h}_{PB_g}$ is far from the target value. In S1, when the sample size increases, although the sampling distribution gets more accurate, the plug-in bandwidths seem to be in general excessively large. In S2, we observe the same pattern, but now the bandwidths become too small for large sample sizes. This biased performance of $\hat{h}_{PB_g}$ may be due, mainly, to two factors. On the one hand, the remaining terms of the bias of (5), depending on $\bar{l}$, do not vanish as fast as required for (15) to be a good bandwidth selector. On the other hand, the method used here (see Reyes et al., 2017) to select the pilot bandwidth, η, requires estimating $A(f')$. This is done by canceling the sum of the two main bias terms of $\hat{\psi}^g_{\eta,2}$. This could be not able to produce good pilot smoothing parameters because, opposite to the complete data case, some second order terms depending on $\bar{l}$ could have a significant impact on the MSE of $\hat{\psi}^g_{\eta,2}$.

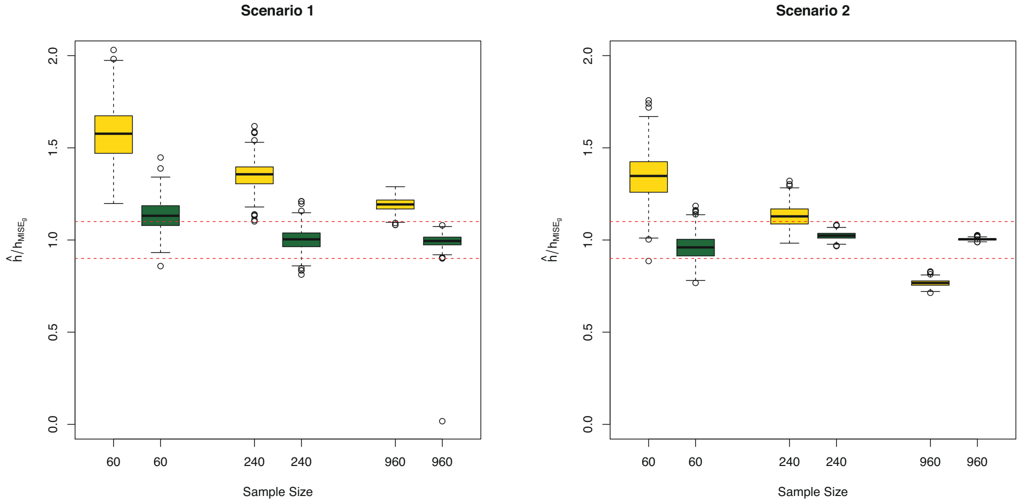

Figure 2: *Box-plots for $\hat{h}_{PB_g}/h_{MISE_g}$ (yellow left box-plots for each sample size) and box-plots for h^*_{MISE}/h_{MISE_g} (green right box-plots for each sample size) for both scenarios. Red dotted lines are plotted at values 0.9 and 1.1 for reference.*

Regarding Figure 2, it can be observed that h^*_{MISE} (green right box-plots for each sample size) outperforms $\hat{h}_{PB_g}$ in approaching h_{MISE_g}. The bootstrap selector shows more stability under any sample size and scenario. This means that it is preferable in both cases of light or heavy grouping and for any sample size.

Figure 3 shows the effect on the distribution estimator (5) of the bandwidth selectors (15) and (17), respectively, in both scenarios. Clearly, when using $\hat{h}_{PB_g}$ (yellow left box-plots for each sample size), while in S1 the quality of $\hat{F}^g_h(x)$ gets better as the sample size increases, in S2, poor distribution estimates for large n are obtained. The impact of poor bandwidth selection is evident in the quality of the distribution estimator, whose error increases by up to three times. However, it should be noted that it does not impact so negatively in the corresponding estimates as in the case of density estimation for grouped data (see Reyes et al., 2017). In opposition, when using the bootstrap bandwidth (green right box-plots for each sample size), the quality of the distribution estimates improves as the sample size increases in both scenarios, clearly outperforming the plug-in bandwidth selector.

It is of interest to study situations in which it is ideally observed the sample size increasing and the average length decreasing at different rates, but in practice that seldom really occurs. For that reason, we also performed some simulations (not shown here for reasons of space, but included in the Supplementary Materials) dealing with a more factual situation in which there is a given sample size and a given set of fixed intervals. In that simulation, a sample size of $n = 240$, a fixed set of average lengths, $\bar{l}$, and a grid of values for h were considered. Those experiments show that the bootstrap smoothing parameter, h^*_{MISE}, seems to be very stable, always centred somewhere around h_{MISE_g}

and with decreasing variability when the average length $\bar{l}$ increases. On the other hand, 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 $\bar{l}$. However, as pointed out previously, it was also observed that bandwidth selection is not so critical in distribution as in density estimation, since slightly different bandwidths produced very similar distribution estimates.

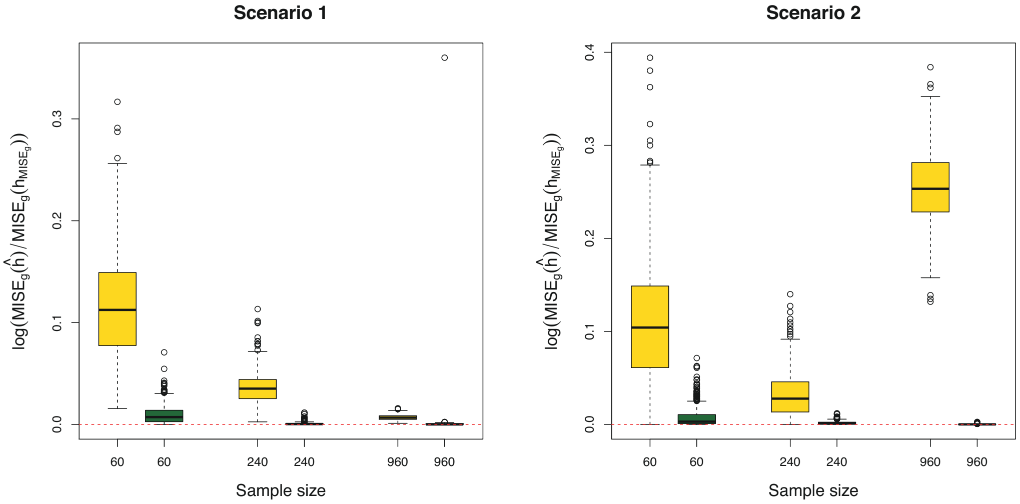

Figure 3: Box-plots for $\ln\left[\mathrm{MISE}_g\left(\hat{h}_{PB_g}\right)/\mathrm{MISE}_g\left(h_{MISE_g}\right)\right]$ *(yellow left box-plots for each sample size) and* $\ln\left[\mathrm{MISE}_g\left(h^*_{MISE}\right)/\mathrm{MISE}_g\left(h_{MISE_g}\right)\right]$ *(green right box-plots for each sample size) for both scenarios.*

6 An empirical study from real data

In this section, a real data set of wild oat (*Avena sterilis* L.) emergence is considered to illustrate the performance of the kernel distribution estimator for grouped data (5), when using the plug-in (15) and bootstrap (17) bandwidth selectors. To do this, the `binnednp` R package (Barreiro et al., 2019) is employed. This package, developed by the authors of the present paper, jointly with a weed scientist and two computer engineers, 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.

The data of *Avena sterilis* were taken from an experiment performed during Winter-Spring 2006-2007 in Gibraleon (37^o 22'N, 6^o 54'W; altitude 26 m), located in the province of Huelva (Andalucia, South of Spain). Four polyvinylchloride cylinders (250 mm diameter 50 mm height) placed 1 m apart were considered and, for each one of them, 200 seeds of *A. sterilis* were mixed thoroughly with the soil and distributed over the 0-100 mm depth. Numbers of emerged weed seedlings were recorded once or twice a week and then removed by cutting seedling stems at ground level with minimum disturbance of the substrate. All the data for the cumulative numbers of seedling emergence from the field were converted to a square meter basis. The CHTT at emergence in the different inspection days, at three depths (10, 20 and 50 mm), were calculated, using the same methodology as that described in Cao et al. (2011).

The observed emergence data are shown in Table 1. As it can be seen, the cumulative hydrothermal time at emergence can not be observed for every individual seed, but just in an aggregated way.

Table 1: *Seedling emergence data of A. sterilis.*

	CHTT			Nº Seedlings				
	Depth			Cylinder				
Date	**10 mm**	**20 mm**	**50 mm**	**1**	**2**	**3**	**4**	**Pooled**
27 November 2006	100	92	67	0	0	0	0	0
4 December 2006	160	146	105	0	0	0	0	0
12 December 2006	218	199	143	2	6	8	3	19
14 December 2006	218	217	155	1	0	0	1	2
19 December 2006	218	217	185	2	1	1	3	7
22 December 2006	218	217	199	2	1	1	0	4
26 December 2006	218	217	204	1	1	0	0	2
28 December 2006	218	217	204	0	0	0	0	0
2 January 2007	218	217	204	0	0	0	0	0
5 January 2007	218	217	204	0	2	0	0	2
9 January 2007	218	217	204	2	2	9	2	15
12 January 2007	218	217	204	3	7	18	11	39
18 January 2007	218	217	204	12	7	19	22	60
25 January 2007	218	217	204	6	5	8	13	32
1 February 2007	265	261	232	2	5	7	7	21
9 February 2007	352	340	287	13	12	5	8	38
15 February 2007	405	421	343	7	12	13	4	36
23 February 2007	459	505	421	0	0	1	0	1
5 March 2007	509	571	538	0	0	0	0	0
19 March 2007	509	571	538	0	0	0	0	0
Nº Emerged seedlings				53	61	90	74	$n = 278$

Before computing the kernel distribution estimator (5) to obtain approximations of the corresponding weed emergence curves, some preliminary analyses were performed. Firstly, the function `anv.binned` included in the `binnednp` package was employed to test whether the "cylinder factor" does not have a significant effect on the emergence curve. If so, we could considered the pooled sample of the four cylinders. In the function `anv.binned`, a bootstrap approach using a Cramér-von Mises type distance is implemented to carry out this type of hypothesis testing. The experimentation conditions seem to support the idea of having a "non-significant cylinder effect" and, after applying the function anv.binned, this hypothesis was corroborated for the three depths. Secondly, an interesting issue for the weed researchers is to find out what the best soil depth is among the three possibilities available in this case, 10, 20 and 50 mm, to measure the CHTT in order to have more prediction power. To address this problem, moment-based indices and probability density-based indices were proposed in Cao et al. (2011). Estimates of these indices are implemented in the function emergence.indices of the binnednp package. After applying this function to the pooled sample of *Avena sterilis*, it is concluded that the best soil depth to measure the CHTT is 10 mm. Therefore, in what follows, only the CHTT measured at 10 mm and the pooled sample are considered for the subsequent analyses.

After these previous analyses, the emergence curve of *Avena sterilis*, using the CHTT measured at 10 mm and the pooled sample, is estimated computing the kernel distribution estimator (5). To do this, we used the functions bw.dist.binned and bw.dist.binned.boot of the binnednp package, returning the plug-in (15) and bootstrap (17) bandwidths, respectively. Arguments in these functions allow to control, among other things, the pilot bandwidths needed in both selectors. For example, in the case of the plug-in bandwidth, $\hat{h}_{PB_g}$, in bw.dist.binned, different types of models can be used in the last step of the iterative method explained in Section 4.1: assuming a normal distribution, using a complete nonparametric approach or considering a normal mixture model. In the case of the bootstrap bandwidth, h^*_{MISE}, in bw.dist.binned.boot, the user can employ as a pilot bandwidth that selected using the method inspired by the idea of smoothing splines, described in Section 4.2, or the one derived using the plug-in technique, $\hat{h}_{PB_g}$. The default pilot bandwidths in these functions are those described in Sections 4.1 and 4.2, respectively. Other parameters of bw.dist.binned and bw.dist.binned.boot allow to plot the corresponding nonparametric distribution estimators and to compute bootstrap confidence bands for the distribution function. It is important to highlight that the functions of this library have been efficiently programmed, using integration of C++ in the R code, and applying parallel computing methods to speed up the running time of the algorithms. This is especially important in those methods making use of bootstrapping to obtain numerical results in a very short time.

Using the default pilot bandwidths, the plug-in and bootstrap smoothing parameters obtained are, respectively, 9.83 and 13.74. The corresponding kernel distribution estimates of the emergence curves computed using (5) are shown in the left panel of Figure 4 (in green when using the plug-in bandwidth and in red when using the bootstrap bandwidth). The empirical distribution of the grouped sample (black line) is also shown in this plot. As indicated in the previous section, it can be seen that the effect of the bandwidth on the estimator's behaviour is not substantial, since slightly different bandwidths produce very similar distribution estimates.

As pointed out in Section 1, parametric regression models have been widely used to model the relationship between the CHTT and weed emergence. For the sake of comparison, the function bw.dist.binned also allows to fit Weibull and logistic parametric regression functions to describe seedling emergence, with parameters estimated by maximum likelihood. The corresponding fits using the *Avena sterilis* data set are shown in the right panel of Figure 4, using a green line for the Weibull and a blue line for the logistic estimators. The nonparametric distribution estimator (5) with bootstrap bandwidth (red line) and the empirical distribution of the grouped sample data (black line) are also included in this plot. It can be observed that none of both classical parametric (distribution) models fits the data well, possibly leading to wrong emergence estimations. On the other hand, the nonparametric approach does not assume any particular distribution for the variable under consideration. As a consequence, it provides more flexible estimators capable of capturing complex features in the HTT distribution and producing more reliable results.

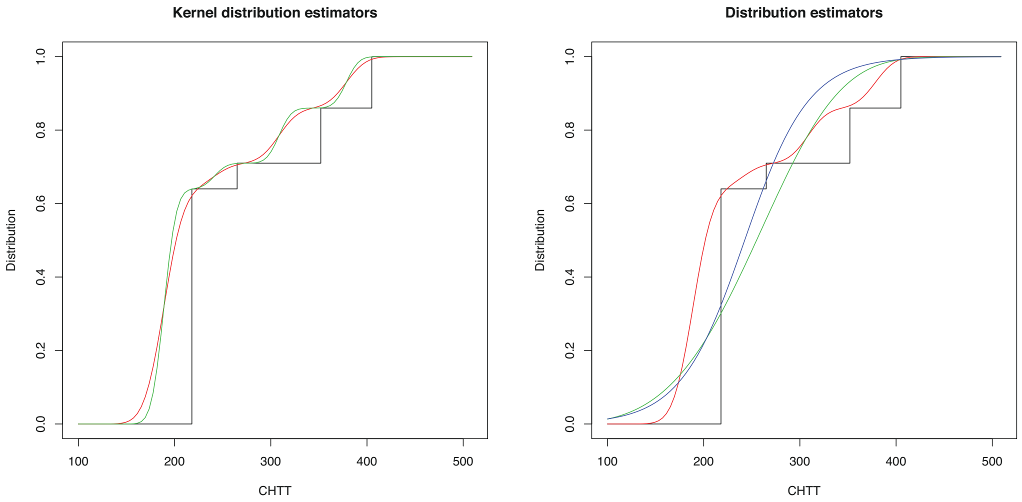

Figure 4: *Left panel: Kernel distribution estimates considering plug-in (green line) and bootstrap (red line) bandwidths. Right panel: parametric regression fits, Weibull (green line) and logistic (blue line), and nonparametric kernel distribution estimate using the bootstrap bandwidth (red line). The empirical distribution of the grouped sample (black lines) is also shown.*

7 Conclusions

In short, it has been shown that under realistic assumptions, the kernel distribution estimator is an effective tool for modeling grouped data due to the good performance of the bootstrap smoothing parameter selector proposed in this paper. This bandwidth selector, using an appropriate criterion to select the corresponding pilot bandwidth, presents a stable and unbiased sampling distribution under any scenario or sample size in the simulation studies performed. Regarding the Polansky and Baker plug-in bandwidth, although theoretically it is a consistent estimator of the optimal bandwidth, in practice, it only has an appropriate behaviour when there is a fixed sample size and a given set of intervals for certain degree of grouping. This can be due to the fact that this plug-in bandwidth is focused on minimizing the AMISE and some neglected terms of less order, depending on $\bar{l}$, can have a substantial influence under certain grouping conditions (see Remark 3.2). Something similar could occur in the process of selecting the pilot bandwidth needed to estimate $A(f')$. On the other hand, h^*_{MISE} targets directly the MISE, producing much better results.

In any case, the different simulations performed show that the kernel distribution estimator is a somewhat robust procedure, in the sense that bandwidth selections slightly different from the optimal bandwidth do not seem to heavily influence the distribution estimation. From another viewpoint, it was shown that really high values of the ratio between the average length $\bar{l}$ and the data range have to be considered in order to actually notice a severe impact of the grouping effect.

These findings leave some insights about kernel distribution estimation for grouped data as well as some possible future work. Since distribution estimation seems to be resistant to grouping effect, a possible future topic of research could be the design of a plug-in bandwidth selector that could work well in different grouping scenarios. This would imply to find out the real influence of second-order terms in the MISE of $\hat{F}_h^g(x)$ and somehow incorporate these effects in the plug-in bandwidth expression. Moreover, a deeper study about the pilot bandwidth selection problem to estimate $A(f')$ would also be necessary. These two issues would transform the usual simple plug-in bandwidth selection method in a much more complicated problem. Fortunately, the bootstrap bandwidth approach proposed in this paper provides a selector that covers any case of grouping, thus controlling or reducing the increase of the error of the estimates. Moreover, it is important to note that this bootstrap procedure does not need Monte Carlo and, therefore, it is also an efficient computing time approach. Facing applications, this implies a substantial improvement in the estimation of data structure, allowing smart inferences even when data are heavily grouped.

Appendix A. Proof of Theorem 3.1

Proof Applying the expectation operator to (5), it is easy to prove that

$$\mathbb{E}\left[\hat{F}_h^g(x)\right] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{E}[w_i] = \sum_{i=1}^{k} \mathbb{K}\left(\frac{x-t_i}{h}\right) p_i \tag{A.1}$$

where $p_i = F(y_i) - F(y_{i-1})$.

Using a Taylor expansion of p_i around t_i and substituting into (A.1), and the fact that

$$\alpha_{ji} = \left(\frac{l_i}{2}\right)^j - \left(-\frac{l_i}{2}\right)^j = \begin{cases} 0 & \text{for j even} \\ 2\left(\frac{l_i}{2}\right)^j & \text{else} \end{cases}, \tag{A.2}$$

gives

$$\mathbb{E}\left[\hat{F}_h^g(x)\right]=\sum_{i=1}^{k} l_i H_1(t_i)+\frac{1}{24}\sum_{i=1}^{k} l_i^3 H_2(t_i)+\frac{1}{4!}\sum_{i=1}^{k}\mathbb{K}\left(\frac{x-t_i}{h}\right)\delta, \tag{A.3}$$

where $\delta=F^{(4)}(\xi_i)\left(\frac{l_i}{2}\right)^4-F^{(4)}(\xi_{i-1})\left(-\frac{l_i}{2}\right)^4$, with $\xi_i\in[t_i,y_i]$ and $\xi_{i-1}\in[y_{i-1},t_i]$, $H_1(t)=F'(t)\,\mathbb{K}\left(\frac{x-t}{h}\right)$ and $H_2(t)=F'''(t)\,\mathbb{K}\left(\frac{x-t}{h}\right)$. As long as $F^{(4)}$ is Lipschitz, δ can be easily bounded leading to $\left|\sum_{i=1}^{k}\mathbb{K}\left(\frac{x-t_i}{h}\right)\delta\right|=O\left(\bar{l}^4\right)$, so that (A.3) becomes

$$\mathbb{E}\left[\hat{F}_h^g(x)\right]=\sum_{i=1}^{k} l_i H_1(t_i)+\frac{1}{24}\sum_{i=1}^{k} l_i^3 H_2(t_i)+O\left(\bar{l}^4\right). \tag{A.4}$$

Considering the first term on the right hand side of (A.4), taking the integral over the i-th interval, using a Taylor expansion with $s=t-t_i$, by (A.2) and summing over all k intervals gives

$$\sum_{i=1}^{k} l_i H_1(t_i)=\int H_1(t)\,dt-\frac{1}{24}\sum_{i=1}^{k} l_i^3 H_1''(t_i)-\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(4)}(\xi_i)\,(t-t_i)^4\,dt. \tag{A.5}$$

Bounding the third term on the right hand side of (A.5) gives

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(4)}(\xi_i)\,(t-t_i)^4\,dt\right|=O\left(\frac{\bar{l}^4}{h^4}\right). \tag{A.6}$$

Now, working on the second term on the right hand side of (A.5), it can be expressed as

$$\sum_{i=1}^{k} l_i^3 H_1''(t_i)=\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right) l_i H_1''(t_i)+\overline{l^2}\sum_{i=1}^{k} l_i H_1''(t_i). \tag{A.7}$$

The second term on the right hand side of (A.7) can be expressed as

$$\overline{l^2}\sum_{i=1}^{k} l_i H_1''(t_i)=$$

$$\overline{l^2}\int H_1''(t)\,dt-\frac{\overline{l^2}}{24}\sum_{i=1}^{k} l_i^3 H_1^{(4)}(t_i)-\frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)\,(t-t_i)^4\,dt. \tag{A.8}$$

Bounding the third and second terms on the right hand side of (A.8) gives

$$\left|\frac{\overline{l^2}}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)\,(t-t_i)^4\,dt\right|=O\left(\frac{\bar{l}^6}{h^6}\right)$$

and $\left|\sum_{i=1}^{k} l_i^3 H_1^{(4)}(t_i)\right| \leqslant O\left(\frac{\bar{l}^4}{h^4}\right)$. In turn, bounding $\overline{l^2}\int H_1''(t)\,dt$ results in

$$\left|\overline{l^2}\int H_1''(t)\,dt\right| \leqslant O\left(\bar{l}^2\right)\int \left|H_1''(t)\right| dt. \tag{A.9}$$

By Assumption 3.1, it is easy to check that $H_1''(t)$ =0 when $\frac{x-t}{h} < -1$, and $H_1''(t) = F'''(t)$ when $\frac{x-t}{h} < -1$. As a consequence,

$$\begin{aligned}\int_{-\infty}^{\infty}\left|H_1''(t)\right| dt &= \int_{-\infty}^{x-h}\left|F'''(t)\right| dt + \int_{x-h}^{x+h}\left|\mathbb{K}\left(\frac{x-t}{h}\right)F'''(t)\right.\\ &- \left.\frac{1}{h}2F''(t)K\left(\frac{x-t}{h}\right) + \frac{1}{h^2}F'(t)K'\left(\frac{x-t}{h}\right)\right| dt. \end{aligned}\tag{A.10}$$

Hence, solving and bounding the right hand side of (A.10) gives

$$\int_{-\infty}^{\infty}\left|H_1''(t)\right| dt = O\left(\frac{1}{h}\right),$$

which implies that

$$\left|\overline{l^2}\int H_1''(t)\,dt\right| = O\left(\frac{\bar{l}^2}{h}\right).$$

Updating (A.7), gives

$$\sum_{i=1}^{k} l_i^3 H_1''(t_i) = \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) + O\left(\frac{\bar{l}^2}{h}\right). \tag{A.11}$$

For bounding the first term on the right hand side of (A.11), realize that by previous elaborations,

$$\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i H_1''(t_i) = \sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right)\int_{y_{i-1}}^{y_i} H_1''(t)\,dt -$$

$$\frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i^3 H_1^{(4)}(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right)\int_{y_{i-1}}^{y_i} H_1^{(6)}(\xi_i)(t-t_i)^4\,dt. \tag{A.12}$$

Under Assumption 3.4, the last two terms can be bounded as

$$\left|\sum_{i=1}^{k}\left(l_i^2 - \overline{l^2}\right) l_i^3 H_1^{(4)}(t_i)\right| \leqslant \max_i \left|l_i^2 - \overline{l^2}\right| k l_{max}^3 \left\|H_1^{(4)}\right\|_{\infty} = o\left(\frac{\bar{l}^4}{h^4}\right) \tag{A.13}$$

and

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1^{(6)}\left(\xi_i\right)\left(t-t_i\right)^4 dt\right| \leqslant$$

$$\frac{1}{4!80}\max_i\left|l_i^2-\overline{l^2}\right|kl_{max}^5\left\|H_1^{(6)}\right\|_\infty = o\left(\frac{\overline{l^6}}{h^6}\right), \qquad \text{(A.14)}$$

and $\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''(t)\,dt$ as

$$\begin{aligned}\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''(t)\,dt\right| &\leqslant \max_i\left|l_i^2-\overline{l^2}\right|\sum_{i=1}^{k}\left|\int_{y_{i-1}}^{y_i}H_1''(t)\,dt\right| \\ &\leqslant o\left(\overline{l^2}\right)\int\left|H_1''(t)\right|dt. \qquad \text{(A.15)}\end{aligned}$$

Using the same arguments as above,

$$\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)\int_{y_{i-1}}^{y_i}H_1''(t)\,dt\right| = o\left(\frac{\overline{l^2}}{h}\right). \qquad \text{(A.16)}$$

Considering (A.12), (A.13), (A.14), (A.15) and (A.16),

$$\sum_{i=1}^{k}l_i^3H_1''(t) = O\left(\frac{\overline{l^2}}{h}\right), \qquad \text{(A.17)}$$

thus leading to

$$\sum_{i=1}^{k}l_iH_1\left(t_i\right) = \int H_1(t)\,dt + O\left(\frac{\overline{l^2}}{h}\right). \qquad \text{(A.18)}$$

Integrating by parts, a change of variable, using a Taylor expansion on F and by kernel properties, lead to

$$\sum_{i=1}^{k}l_iH_1\left(t_i\right) = F(x) + \frac{h^2}{2}F''(x)\mu_2(K) + O\left(h^4\right) + O\left(\frac{\overline{l^2}}{h}\right). \qquad \text{(A.19)}$$

Regarding the second term on the right hand side of (A.4),

$$\sum_{i=1}^{k}l_i^3H_2\left(t_i\right) = \sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_iH_2\left(t_i\right) + \overline{l^2}\sum_{i=1}^{k}l_iH_2\left(t_i\right). \qquad \text{(A.20)}$$

Proceeding as above,

$$\left|\sum_{i=1}^{k}\left(l_i^2-\overline{l^2}\right)l_iH_2\left(t_i\right)\right|=o\left(\overline{l}^2\right). \tag{A.21}$$

As to the second term, note that by Ostrowski's inequality (Anastasiou, Kechriniotis, and Kotsos, 2006; Ostrowski, 1938)

$$\left|l_iH_2\left(t_i\right)-\int_{y_{i-1}}^{y_i}H_2\left(t\right)dt\right|\leqslant\frac{1}{4}\mathfrak{L}_{H_2}l_i^2,$$

where $\mathfrak{L}_{H_2}$ is the H_2 Lipschitz constant. Summing up over all k intervals and considering Assumption (3.4) lead to

$$\sum_{i=1}^{k}\left|l_iH_2\left(t_i\right)-\int_{y_{i-1}}^{y_i}H_2\left(t\right)dt\right|=O\left(\frac{\overline{l}}{h}\right)$$

which in turn implies that

$$\overline{l^2}\sum_{i=1}^{k}l_iH_2\left(t_i\right)=\overline{l}^2\int H_2\left(t\right)dt+o\left(\overline{l}^2\right).$$

Integrating by parts, a change of variable, by a Taylor expansion on F'' and simplifying due to the kernel K properties lead to

$$\int H_2\left(t\right)dt=F''\left(x\right)+\frac{h^2}{2}F^{(4)}\left(x\right)\mu_2\left(K\right)+O\left(h^3\right),$$

so that

$$\overline{l^2}\sum_{i=1}^{k}l_iH_2\left(t_i\right)=\overline{l}^2\left[F''\left(x\right)+\frac{h^2}{2}F^{(4)}\left(x\right)\mu_2\left(K\right)+O\left(h^3\right)\right]+o\left(\overline{l}^2\right). \tag{A.22}$$

Using (A.22) and (A.21), Eq. (A.20) becomes

$$\sum_{i=1}^{k}l_i^3H_2\left(t_i\right)=\overline{l}^2\left[F''\left(x\right)+\frac{h^2}{2}F^{(4)}\left(x\right)\mu_2\left(K\right)+O\left(h^3\right)\right]+o\left(\overline{l}^2\right). \tag{A.23}$$

So, joining (A.23) and (A.19) into (A.3), and by Assumption 3.4,

$$\mathbb{E}\left[\hat{F}_h^g\left(x\right)\right]=F\left(x\right)+\frac{h^2}{2}F''\left(x\right)\mu_2\left(K\right)+o\left(h^2\right),$$

from which, the bias is

$$\mathbb{B}\left[\hat{F}_h^g(x)\right] = \frac{1}{2}h^2 F''(x)\,\mu_2(K) + o\left(h^2\right). \tag{A.24}$$

Regarding the variance, considering that $(n_1, n_2, \ldots, n_k)$ is a multinomial random vector and $w_i = n_i/n$, applying this operator to (5), it gives

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i\left(1-p_i\right) - \frac{2}{n}\sum_{i<j}\mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)p_i p_j. \tag{A.25}$$

Since $p_i = F(y_i) - F(y_{i-1})$, using Taylor expansions around t_i and by (A.2), the first term on the right hand side of (A.25) (except a factor $1/n$) can be written as

$$\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i\left(1-p_i\right) = \sum_{i=1}^{k} l_i H_3(t_i) + O\left(\bar{l}\right), \tag{A.26}$$

where $H_3(t) = \mathbb{K}^2\left(\frac{x-t}{h}\right)F'(t)$. Integrating H_3 over the i-th interval, by a Taylor expansion, using $s = t - t_i$, by parity conditions (A.2), summing over all k intervals and reordering gives

$$\sum_{i=1}^{k} l_i H_3(t_i) = \int H_3(t)\,dt - \frac{1}{24}\sum_{i=1}^{k} l_i^3 H_3''(t_i) - \frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_3^{(4)}(\xi_i)\left(t-t_i\right)^4 dt. \tag{A.27}$$

As done for (A.5), it is easy to check that

$$\left|\frac{1}{4!}\sum_{i=1}^{k}\int_{y_{i-1}}^{y_i} H_3^{(4)}(\xi_i)\left(t-t_i\right)^4 dt\right| = O\left(\frac{\bar{l}^4}{h^4}\right) \tag{A.28}$$

and

$$\sum_{i=1}^{k} l_i^3 H_3''(t_i) = O\left(\frac{\bar{l}^2}{h}\right). \tag{A.29}$$

Considering (A.29), (A.28) and (A.27), Eq. (A.26) transforms into

$$\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i\left(1-p_i\right) = \int H_3(t)\,dt + O\left(\bar{l}\right). \tag{A.30}$$

As above, using integration by parts, the change of variable $u = (x-t)/h$ and a Taylor expansion give

$$\int H_3(t)\,dt = F(x) - hF'(x)C_0 + O\left(h^2\right),$$

where $C_0 = 2\int \mathbb{K}(u)K(u)udu$. Substituting the last expression into (A.30) and by Assumption 3.4 it gives

$$\sum_{i=1}^{k}\mathbb{K}^2\left(\frac{x-t_i}{h}\right)p_i(1-p_i) = F(x) - hF'(x)C_0 + O\left(h^2\right). \tag{A.31}$$

Let us turn back to eq. (A.25). Because $p_i = F(y_i) - F(y_{i-1})$, using Taylor expansions around t_i, by (A.2), the second term on the right hand side of (A.25) (except a factor $-2/n$) can be written as

$$\sum_{i<j}\mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)p_ip_j = \sum_{i<j}H_4(t_i,t_j)l_il_j + O\left(\bar{l}^2\right), \tag{A.32}$$

where $H_4(z_1,z_2) = \mathbb{K}\left(\frac{x-z_1}{h}\right)\mathbb{K}\left(\frac{x-z_2}{h}\right)F'(z_1)F'(z_2)$.

Considering the second order Taylor expansion around (t_i,t_j) and by parity conditions (A.2),

$$\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j} H_4(z_1,z_2)\,dz_2dz_1 \;=\; H_4(t_i,t_j)l_il_j + \frac{\mathfrak{T}_0}{2}, \tag{A.33}$$

where

$$\begin{aligned}\mathfrak{T}_0 \;=\; & \int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}\left[\frac{\partial^2 H_4}{\partial z_1^2}(\xi_1,\xi_2)(z_1-t_i)^2 + 2\frac{\partial^2 H_4}{\partial z_1\partial z_2}(\xi_1,\xi_2)(z_1-t_i)(z_2-t_j)\right.\\ & \left.+\frac{\partial^2 H_4}{\partial z_2^2}(\xi_1,\xi_2)(z_2-t_j)^2\right]dz_2dz_1.\end{aligned}$$

Summing over all $k(k-1)/2$ terms of the form (A.33) and reordering,

$$\sum_{i<j}l_il_jH_4(t_i,t_j) \;=\; \sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}H_4(z_1,z_2)\,dz_2dz_1 - \frac{1}{2}\sum_{i<j}\mathfrak{T}_0. \tag{A.34}$$

The second term on the right hand side of (A.34) can be easily bounded by Assumption 3.4, since $\left|\frac{1}{2}\sum_{i<j}\mathfrak{T}_0\right| = O\left(\frac{\bar{l}^2}{h^2}\right)$.

As a consequence,

$$\sum_{i<j}l_il_jH_4(t_i,t_j) = \sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}H_4(z_1,z_2)\,dz_2dz_1 + O\left(\frac{\bar{l}^2}{h^2}\right). \tag{A.35}$$

On the other hand, it is straightforward to prove that

$$\sum_{i<j}\int_{y_{i-1}}^{y_i}\int_{y_{j-1}}^{y_j}H_4(z_1,z_2)\,dz_2dz_1 = \frac{1}{2}\int\int H_4(z_1,z_2)\,dz_2dz_1 + O\left(\bar{l}\right). \tag{A.36}$$

Now, using (A.36) and (A.35),

$$\sum_{i<j} l_i l_j H_4(t_i,t_j) = \frac{1}{2}\int\int H_4(z_1,z_2)\,dz_2dz_1 + O\left(\bar{l}\right) + O\left(\frac{\bar{l}^2}{h^2}\right). \tag{A.37}$$

Integration by parts, two changes of variable $[u_1 = (x - z_1)/h,\ u_2 = (x - z_2)/h]$ and a Taylor expansion around x give

$$\frac{1}{2}\int\int H_4(z_1,z_2)\,dz_2dz_1 = \frac{1}{2}\left[F^2(x) + O\left(h^2\right)\right] \tag{A.38}$$

so that, considering (A.38), (A.37) and Assumption 3.4, Eq. (A.32) becomes

$$\sum_{i<j} \mathbb{K}\left(\frac{x-t_i}{h}\right)\mathbb{K}\left(\frac{x-t_j}{h}\right)p_ip_j = \frac{1}{2}F^2(x) + O\left(h^2\right). \tag{A.39}$$

Now, putting back (A.39) and (A.31) in (A.25) and simplifying,

$$\mathbb{V}\left[\hat{F}_h^g(x)\right] = \frac{1}{n}F(x)\left[1-F(x)\right] - \frac{h}{n}F'(x)C_0 + O\left(\frac{h^2}{n}\right). \tag{A.40}$$

Collecting (A.40) and (A.24), one obtains

$$\begin{aligned}\text{MSE}\left[\hat{F}_h^g(x)\right] \;=\; & \frac{1}{4}h^4F''(x)^2\mu_2(K)^2 + \frac{1}{n}F(x)\left[1-F(x)\right] - \frac{h}{n}F'(x)C_0 \\ & + O\left(\frac{h^2}{n}\right) + o\left(h^4\right).\end{aligned} \tag{A.41}$$

Finally, dealing with the integrated versions of the terms coming up in the proof of (A.41), one can obtain the following asymptotic expression for AMISE,

$$\text{AMISE}\left[\hat{F}_h^g\right] = \frac{1}{4}h^4\mu_2(K)^2A(f') + \frac{1}{n}\int F(x)\left[1-F(x)\right]dx - \frac{h}{n}C_0,$$

which corresponds with just integrating the leading terms in (A.41). ■

Acknowledgements

The authors thank three anonymous referees and the Editor for numerous useful comments that significantly improved this article. The authors also thank Dr. Fernando Bastida and Dr. José Luis González-Andújar for providing the *Avena sterilis* L. emergence data employed in Section 6.

This research has been supported by MINECO grants MTM2014-52876-R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva

ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

Altman, N. and Leger, C. (1995). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference*, 46, 195–214.

Anastasiou, K., Kechriniotis A. and Kotsos, B. (2006). Generalizations of the Ostrowski's inequality. *Journal of Interdisciplinary Mathematics*, 9, 49–60.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68, 326–328.

Barreiro, D., Fraguela, B., Doallo, R., Cao, R., Francisco-Fernandez, M. and Reyes, M. (2019). *binnednp: Nonparametric estimation for interval-grouped data*. https://cran.r-project.org/package=binnednp R package version 0.4.0.

Blower, G. and Kelsall, J. E. (2002). Nonlinear kernel density estimation for binned data: convergence in entropy. *Bernoulli*, 8, 423–449.

Bowman, A., Hall, P. and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85, 799–808.

Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via waved block thresholding. *Probability Theory and Related Fields*, 146, 401–433.

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F. and González-Andújar, J. L. (2011). Computing statistical indices for hydrothermal times using weed emergence data. *Journal of Agricultural Science*, 149, 701–712.

Cao, R., Francisco-Fernández, M., Anand, A., Bastida, F. and González-Andújar, J. L. (2013). Modeling *Bromus diandrus* seedling emergence using nonparametric estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 64–86.

Coit, D. and Dey, K. (1999). Analysis of grouped data from field-failure reporting systems. *Reliability Engineering & System Safety*, 65, 95–101.

Dutta, S. (2015). Local smoothing for kernel distribution function estimation. *Communications in Statistics, Simulation and Computation*, 44, 878–891.

González-Andújar, J. L., Francisco-Fernández, M., Cao, R., Reyes, M., Urbano, J. M., Forcella, F. and Bastida, F. (2016). A comparative study between nonlinear regression and nonparametric approaches for modeling *Phalaris paradoxa* seedling emergence. *Weed Research*, 56, 367–376.

Guo, S. (2005). Analysing grouped data with hierarchical linear modeling. *Children and Youth Services Review*, 27, 637–652.

Hill, P. (1985). Kernel estimation of a distribution function. *Communications in Statistics, Theory and Methods*, 14, 605–620.

Klein, J. P. and Moeschberger, M. (1997). *Survival Analysis*. New York: Springer Verlag.

Mächler, M. (2017). *nor1mix: Normal (1-d) Mixture Models (S3 Classes and Methods)*. https://CRAN.R-project.org/package=nor1mix. R package version 1.2-3.

Mack, Y. (1984). Remarks on some smoothed empirical distribution functions and processes. *Bulletin of Informatics and Cybernetics*, 21, 29–35.

Mclachlan, G. and Peel, D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.

Minoiu, C. and Reddy, S. (2009). Estimating poverty and inequality from grouped data: How well do parametric methods perform? *Journal of Income Distribution*, 18, 160–178.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Applications*, 10, 186–190.

Ostrowski, A. (1938). Über die Absolutabweichung einer differenzierbaren Funktion von ihrem Integralmittelwert. *Commentarii Mathematici Helvetici*, 10, 226–227.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, 1065–1076.

Pipper, C. and Ritz, C. (2007). Checking the grouped data version of the Cox model for interval-grouped data survival data. *Scandinavian Journal of Statistics*, 34, 405–418.

Polanski, A. and Baker, E. (2000). Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, 65, 63–80.

Quintela-del-Río, A. and Estévez-Pérez, G. (2012). Nonparametric kernel distribution function estimation with kerdiest: An R package for bandwidth choice and applications. *Journal of Statistical Software*, 50, 1–21. http://www.jstatsoft.org/v50/i08/.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reiss, R. (1981). Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, 8, 116–119.

Reyes, M., Francisco-Fernandez, M. and Cao, R. (2016). Nonparametric kernel density estimation for general grouped data. *Journal of Nonparametric Statistics*, 28, 235–249.

Reyes, M., Francisco-Fernández, M. and Cao, R. (2017). Bandwidth selection in kernel density estimation for interval-grouped data. *TEST*, 26, 527–545.

Rizzi, S., Thinggaard, M., Engholm, G., Christensen, N., Johannesen, T., Vaupel, J. and Jacobsen, R. (2016). Comparison of non-parametric methods for ungrouping coarsely aggregated data. *BMC Medical Research Methodology*, 16, 59.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27, 832–837.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution function. *Journal of Statistical Planning and Inference*, 35, 65–75.

Scott, D. and Sheather, S. (1985). Kernel density estimation with binned data. *Communications in Statistics, Theory and Methods*, 27, 832–837.

Titterington, D. (1983). Kernel-based density estimation using censored, truncated or grouped data. *Communications in Statistics, Theory and Methods*, 12, 2151–2167.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38, 290–295.

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

Wang, B. and Wang, X.-F. (2016). Fitting the generalized lambda distribution to pre-binned data. *Journal of Statistical Computation and Simulation*, 86, 1785–1797.

Wang, B. and Wertelecki, W. (2013). Density estimation for data with rounding errors. *Computational Statistics & Data Analysis*, 65, 4–12.