Polygonal smoothing of the empirical distribution function

Delphine Blanke, Denis Bosq

POLYGONAL SMOOTHING OF THE EMPIRICAL DISTRIBUTION FUNCTION

DELPHINE BLANKE AND DENIS BOSQ

Abstract. We present two families of polygonal estimators of the distribution function: the first family is based on the knowledge of the support while the second addresses the case of an unknown support. Polygonal smoothing is a simple and natural method for regularizing the empirical distribution function F_n but its properties have not been studied deeply. First, consistency and exponential type inequalities are derived from well-known convergence properties of F_n. Then, we study their mean integrated squared error (MISE) and we establish that polygonal estimators may improve the MISE of F_n. We conclude by some numerical results to compare these estimators globally, and also together with the integrated kernel distribution estimator. Distribution function estimation, Polygonal estimator, Cumulative frequency polygon, Order statistics, Mean integrated squared error, Exponential inequality, Smoothed processes 62G05, 62G30, 62G20

1. Introduction

Distribution function estimation finds many applications especially in survival analysis or for quantile estimation of a population. This explain why a large number of articles are devoted to its estimation, see e.g. the nice review written by Servien (2009). The most classical and simple estimate of the cumulative distribution function (cdf) F is the empirical distribution function (edf) F_n. Properties of F_n are well-known, it is an unbiased and strongly uniformly consistent estimator of F. But if F is absolutely continuous with density f, the edf F_n does not take into account this smoothness property. Several smoothing techniques have been considered in the literature to correct this drawback. The kernel estimator is one of the most commonly used. Theoretical properties of this estimator are now well established, we may refer to Swanepoel and Van Graan (2005); Servien (2009) and more recently to Quintela-del Río and Estévez-Pérez (2012) for a literature review about it. Other techniques are possible such as moving polynomial regression (Lejeune and Sarda, 1992), polynomial spline regression (Cheng and Peng, 2002), level crossings (Huang and Brill, 2004), Bernstein polynomials (with degree depending on the sample size n, Babu et al., 2002; Leblanc, 2012) among others. All these methods depend on some smoothing parameter to calibrate to avoid classical phenomena of over or under-smoothing.

For density estimation, polygonal frequency estimators, connecting mid-bin values of an histogram by straight lines, are classical and have been widely studied, see eg Scott (1985), Simonoff (1996, p. 20-39) and references therein. For estimating the distribution function, practitioners are also accustomed to use the cumulative frequency polygons (hand drawing graphs, sometimes called ogives), especially for grouped data. In this paper, we consider n independent and identically distributed (iid) real-valued $X_1, \ldots, X_n$ with an absolutely continuous cumulative distribution function (cdf) F having the density f. For estimating F, a simple polygonal smoothing of the empirical distribution function is considered. More precisely, we present and study two families of polygonal estimators of F: the first one supposes

that the support $[a, b]$ of f is known while the second one addresses the case of unknown support. Until now, and as far as we can judge, comparison of these simple estimators with the classical edf has not been studied deeply. To our knowledge, the single reference is Read (1972) where a continuous estimator of F, namely joining the points $(X_i^*, \frac{i}{n+1})$, with $X_1^*, \ldots, X_n^*$ the ordered sample, is studied. For n sufficiently large, it is shown that his expected squared error is no larger than that of F_n and that it dominates F_n in terms of the integrated error (with respect to F). Our goal is to extend these first results to our two general families of smoothed estimators of F_n.

The paper is organized as follows. In Section 2, we introduce our two families of polygonal estimators and give their first properties that can be deduced from their proximity to F_n. In Section 3, the MISE of these estimators is established in Theorem 3.6, the main result of this paper. To examine the small sample behaviour of the polygonal estimators, we present in Section 4 a simulation study that includes also the kernel distribution estimator. To this aim, we consider the mixtures of normal distributions introduced by Marron and Wand (1992) and completed by Janssen et al. (1995). A conclusion and discussion about the possible extensions of our results appear in Section 5. Finally, the most technical auxiliary result is proved in the Appendix A and parameters of the normal mixtures involved in the simulations are recalled in the Appendix B.

2. Polygonal distribution estimators

2.1. **Definition.** For i.i.d random variables $X_1, \ldots, X_n$ with absolutely continuous cdf F, we consider two families of polygonal estimators of F derived from the classical empirical distribution function F_n:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i),\ x \in \mathbb{R}$$

where $\mathbb{I}_A$ denotes the indicator function of the set A. The first family adresses the case of a known support $[a, b]$ while the second is adapted to the case of unknown/infinite support. If $X_1^* < \cdots < X_n^*$ (almost surely) denotes the ordered sample, we define $G_n^{(j,p)}$, $j = 1, 2$ as follows (see also Figure 1 for an illustration with $p = 0$, $p = \frac{1}{2}$ and $p = 1$):

$$G_n^{(1,p)}(t) = \frac{(t-a)(1-p)}{n(X_1^* - a)} \mathbb{I}_{[a,X_1^*)}(t) + \big(1 - \frac{p(b-t)}{n(b-X_n^*)}\big) \mathbb{I}_{[X_n^*,b]}(t) + \sum_{k=1}^{n-1} \frac{t + (k-p)X_{k+1}^* - (k+1-p)X_k^*}{n(X_{k+1}^* - X_k^*)} \mathbb{I}_{[X_k^*, X_{k+1}^*)}(t) \quad (1)$$

and

$$G_n^{(2,p)}(t) = G_n^{(1,p)}(t)\, \mathbb{I}_{[X_1^*, X_n^*)}(t) + \max\big(0, \frac{t - (2-p)X_1^* + (1-p)X_2^*}{n(X_2^* - X_1^*)}\big) \mathbb{I}_{(-\infty, X_1^*)}(t) + \min(1, \frac{t + (n-1-p)X_n^* - (n-p)X_{n-1}^*}{n(X_n^* - X_{n-1}^*)}) \mathbb{I}_{[X_n^*, +\infty)}(t). \quad (2)$$

By this way, their construction depend on a known real parameter p, chosen in [0,1], indicating that the sample points $(X_k^* + p(X_{k+1}^* - X_k^*), \frac{k}{n})$, $k = 1, \ldots, n-1$, are connected. For example, the case $p = 0$ (respectively $p = 1$) joins the points $(X_k^*, \frac{k}{n})$ (resp. $(X_{k+1}^*, \frac{k}{n})$) while the midpoints $(\frac{X_k^* + X_{k+1}^*}{2}, \frac{k}{n})$ are connected when

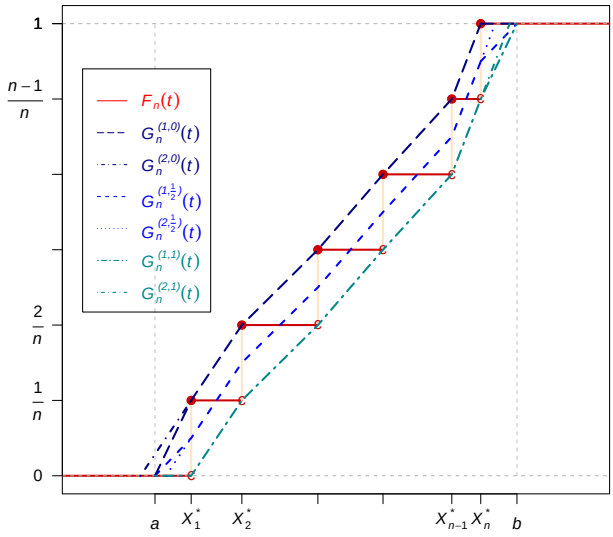

FIGURE 1. Estimators F_n, $G_n^{(j,0)}$, $G_n^{(j,\frac{1}{2})}$, $G_n^{(j,1)}$, $j = 1, 2$, for a distribution with support $[a, b]$

$p = \frac{1}{2}$. These two families of estimators constitute proper distribution functions as continuous and increasing functions.

Note that Read (1972) has studied an estimator very similar to $G_n^{(1,0)}$ (namely, with $a = 0$, $b = 1$ and $n + 1$ at the denominator rater than n). He established that its pointwise risk is no larger than that of F_n and, that it dominates F_n in terms of integrated risk (however, this latter result is stated without proof). In the following, we study the exact second order asymptotic behaviour of the MISE of our general estimators.

Remark 2.1. *The families $G_n^{(1,p)}$ and $G_n^{(2,p)}$ differ only at the two ends of the interval $[a, b]$: the first uses the support in its construction while the second one is adapted to the case of unknown or unbounded support. We may remark as in Babu et al. (2002) that to obtain the bounded support $[0, 1]$, a monotone transformation like $Y = X/(1 + X)$ may handle the case of random variables with support $[0; \infty)$, and $Y = (1/2) + (\tan^{-1} X/\pi)$ can be taken for real random variables. Then estimators of the cdf may be derived with $\widetilde{F}_n^{(p)}(x) = G_n^{(1,p)}(y)$ (and $[a, b] = [0, 1]$). Also, in our numerical studies of Section 4 where random variables are real-valued, $[a, b]$ is taken as a prediction interval with quite good results.*

2.2. **Properties.** First, remark that $G_n^{(1,p)}$ and $G_n^{(2,p)}$ are equal to $\frac{k-p}{n}$ at sample points. So for $k = 1, \ldots, n-1$, $j = 1, 2$:

$$G_n^{(j,p)}(X_{k+1}^*) - G_n^{(j,p)}(X_k^*) = F_n(X_{k+1}^*) - F_n(X_k^*) = \frac{1}{n}.$$

Moreover, it is easy to show that for p in $[0, 1)$ and $j = 1, 2$, $G_n^{(j,p)}(t)$ and $F_n(t)$ are equal for $t = X_k^* + p(X_{k+1}^* - X_k^*)$, $k = 1, \ldots, n-1$. For example, $G_n^{(j,\frac{1}{2})}$ intersects F_n in the middle of $[X_k^*, X_{k+1}^*]$, $k = 1, \ldots, n-1$. The case $p = 1$ is specific with $G_n^{(j,1)}(X_{k+1}^*) = F_n(X_k^*)$, $j = 1, 2$, $k = 1, \ldots, n-1$.

For random variables with support $[a,b]$, one has also $G_n^{(1,p)}(a) = F_n(a) = 0$ and $G_n^{(1,p)}(b) = F_n(b) = 1$. On the other hand, we may note that

$$G_n^{(2,p)}(t) = \begin{cases} 0 & \text{for } t \leq (2-p)X_1^* - (1-p)X_2^* \leq X_1^* \\ 1 & \text{for } t \geq (1+p)X_n^* - pX_{n-1}^* \geq X_n^* \end{cases}$$

but $\Big[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\Big]$ is not necessarily included in $[a,b]$.

With the help of known properties of F_n, it is easy to obtain first results of convergence. To simplify the presentation of our results and without loss of generality, we suppose from now on that $[a,b] = [0,1]$ in the definition of the family $G_n^{(1,p)}$. We start with a lemma that gives the proximity of the estimators $G_n^{(j,p)}$ with F_n. This lemma is simple but essential for all the results derived in this paper.

Lemma 2.2. *We get*

(a)

$$G_n^{(1,p)}(t) - F_n(t) = \frac{(1-p)t}{nX_1^*}\,\mathbb{I}_{[0,X_1^*)}(t) - \frac{(1-t)p}{n(1-X_n^*)}\,\mathbb{I}_{[X_n^*,1]}(t) + \sum_{k=1}^{n-1} \frac{t - pX_{k+1}^* - (1-p)X_k^*}{n(X_{k+1}^* - X_k^*)}\,\mathbb{I}_{[X_k^*, X_{k+1}^*)}(t); \quad (3)$$

(b)

$$G_n^{(2,p)}(t) - F_n(t) = \frac{t + (1-p)X_2^* - (2-p)X_1^*}{n(X_2^* - X_1^*)}\,\mathbb{I}_{[(2-p)X_1^* - (1-p)X_2^*, X_1^*)}(t) + \frac{t - (1+p)X_n^* + pX_{n-1}^*}{n(X_n^* - X_{n-1}^*)}\,\mathbb{I}_{[X_n^*, (1+p)X_n^* - pX_{n-1}^*]}(t) + \sum_{k=1}^{n-1} \frac{t - pX_{k+1}^* - (1-p)X_k^*}{n(X_{k+1}^* - X_k^*)}\,\mathbb{I}_{[X_k^*, X_{k+1}^*)}(t). \quad (4)$$

Proof. Straightforward from the definition of F_n and relations (1)-(2). □ □

From Lemma 2.2, one may deduce that

$$\frac{1}{2n} \leq \Big\|F_n - G_n^{(j,p)}\Big\|_\infty = \max\big(\frac{p}{n}, \frac{1-p}{n}\big) \leq \frac{1}{n} \quad (5)$$

so that $\Big\|F_n - G_n^{(j,p)}\Big\|_\infty$ is maximal and equals to $\frac{1}{n}$ for $p = 0$ or $p = 1$. Its minimal value is reached for $p = \frac{1}{2}$ with $\frac{1}{2n}$. Next, we may derive an exponential inequality.

Proposition 2.3. *We obtain*

(a) For $j = 1, 2$ and $\varepsilon > \frac{1}{2n(1-a_0)}$ with $0 < a_0 < 1$,

$$\mathbb{P}\big(\Big\|G_n^{(j,\frac{1}{2})} - F\Big\|_\infty \geq \varepsilon\big) \leq 2\exp(-2a_0^2 n\varepsilon^2),\ n \geq 1.$$

(b) More generally,

$$\mathbb{P}\big(\Big\|G_n^{(j,p)} - F\Big\|_\infty \geq \varepsilon\big) \leq 2\exp(-2a_0^2 n\varepsilon^2),\ 0 < a_0 < 1,$$

for $\varepsilon > \frac{\max(p,1-p)}{n(1-a_0)}$, $n \geq 1$.

Proof. First, for all $p \in [0,1]$, we have

$$G_n^{(j,p)} - F = \big(G_n^{(j,p)} - F_n\big) + \big(F_n - F\big)$$

so by (5)

$$\Big\|G_n^{(j,p)} - F\Big\|_\infty \le \max\big(\frac{p}{n}, \frac{1-p}{n}\big) + \|F_n - F\|_\infty .$$

Thus, since F is supposed to be absolutely continuous in our framework, Massart's inequality (1990) gives

$$\begin{aligned}\mathbb{P}\big(\Big\|G_n^{(j,p)} - F\Big\|_\infty \ge \varepsilon\big) &\le \mathbb{P}\Big(\|F_n - F\|_\infty \ge \varepsilon - \max\big(\frac{p}{n}, \frac{1-p}{n}\big)\Big)\\ &\le 2\exp\Big(-2n\big(\varepsilon - \max\big(\frac{p}{n}, \frac{1-p}{n}\big)\big)^2\Big).\end{aligned}$$

Next, (b) holds with $\varepsilon - \max\big(\frac{p}{n}, \frac{1-p}{n}\big) \ge a_0\varepsilon$ for $\varepsilon > \frac{\max(p,1-p)}{n(1-a_0)}$ and one derives (a) with the choice $p = \frac{1}{2}$. □ □

Another derivation is envisaged in the following proposition.

Proposition 2.4. *We get*
(a) for $p = \frac{1}{2}$, $j = 1, 2$:

$$\mathbb{P}\big(\Big\|G_n^{(j,\frac{1}{2})} - F\Big\|_\infty \ge \varepsilon\big) \le 2\exp(-2n\varepsilon^2),\ 0 < \varepsilon < \frac{1}{4n},\ n \ge 1;$$

(b) and more generally,

$$\mathbb{P}\big(\Big\|G_n^{(j,p)} - F\Big\|_\infty \ge \varepsilon\big) \le 2\exp(-2n\varepsilon^2),\ 0 < \varepsilon < \max\big(\frac{p}{2n}, \frac{1-p}{2n}\big),$$

$n \ge 1$.

Proof. For all $p \in [0,1]$, we have again:

$$\begin{aligned}\mathbb{P}\big(\Big\|G_n^{(j,p)} - F\Big\|_\infty \ge \varepsilon\big) &\le 2\exp\Big(-2n\Big(\varepsilon - \max\big(\frac{p}{n}, \frac{1-p}{n}\big)\Big)^2\Big)\\ &= 2\exp(-2n\varepsilon^2)\exp\Big(-\frac{2}{n}\big(\max(p,1-p)\big)^2 + 4\varepsilon\max(p,1-p)\Big).\end{aligned}$$

So, for $n \ge 1$, one gets the result $\mathbb{P}\big(\Big\|G_n^{(j,p)} - F\Big\|_\infty \ge \varepsilon\big) \le 2\exp(-2n\varepsilon^2)$ as

$$2n\varepsilon < \max(p,1-p) \Longleftrightarrow 4\varepsilon\max(p,1-p) - \frac{2}{n}\big(\max(p,1-p)\big)^2 < 0.$$

□ □

Finally, note that known asymptotic results on F_n allow to derive easily limits in distribution for the estimators $G_n^{(j,p)}$, $j = 1, 2$ and $p \in [0,1]$. From (5), we get

$$\sup_t \big|\sqrt{n}\big(G_n^{j,p}(t) - F(t)\big) - \sqrt{n}\big(F_n(t) - F(t)\big)\big| \le \frac{1}{\sqrt{n}}$$

and Theorem 3.1 in Billingsley (1999, p.27) implies that $G_n^{(j,p)}$ and F_n, properly normalized, have the same limit in distribution.

3. Mean integrated error of polygonal estimators

To establish the mean integrated error of the polygonal estimators, we suppose from now on that $X_1, \ldots, X_n$, n are i.i.d. random variables with absolutely continuous cdf F such that f has compact support $[0,1]$.

3.1. **Asymptotic integrated bias.** We begin with a technical lemma useful for further derivations of our results.

Lemma 3.1. *For all $m \in \mathbb{N}^*$ and $p \in [0,1]$,*

(a)

$$\int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right)^m \mathrm{d}t = \frac{\left((1-p)^m - (-1)^m p^m\right)\left(pX_1^* + (1-p)X_n^*\right) + (-1)^m p^m}{(m+1)n^m};$$

(b)

$$\int_{-\infty}^{\infty} \left(G_n^{(2,p)}(t) - F_n(t)\right)^m \mathrm{d}t = \frac{(1-p)^{m+1}X_2^* - X_1^*\left(2(1-p)^{m+1} + (-1)^m p^{m+1}\right)}{(m+1)n^m} + \frac{X_n^*\left(2(-1)^m p^{m+1} + (1-p)^{m+1}\right) + X_{n-1}^*(-1)^{m+1}p^{m+1}}{(m+1)n^m}.$$

Proof. Straightforward from Lemma 2.2 and formulas (3)-(4) raised to the power of m and integrated term by term. □ □

Note that for $p = \frac{1}{2}$ and m even, $\int_0^1 \left(G_n^{(1,\frac{1}{2})}(t) - F_n(t)\right)^m \mathrm{d}t$ becomes constant and equal to $\frac{(2n)^{-m}}{m+1}$.

In the sequel, we will use the following conditions on the density f.

Assumption 3.1 (A3.1)**.**

(i) f is continuous on [0,1] and $\inf_{x\in[0,1]} f(x) \geq c_0$ for some positive constant c_0;

(ii) f is a Lipschitz function: there exists a positive constant c_1 such that for all $(x,y) \in (0,1)^2$, $|f(x) - f(y)| \leq c_1 |x - y|$.

Note that A3.1-(ii) is less stringent than usual conditions for kernel distribution estimators where, in general, f is supposed to be at least differentiable, see e.g. Azzalini (1981); Swanepoel (1988); Jones (1990). The condition of minoration in A3.1-(i) is of course more stringent but it is useful to derive the following lemma where equivalent expressions are obtained for expectations of functions of $X_1^*, \ldots, X_n^*$.

Lemma 3.2. *If the condition A3.1-(i) holds then, for all integers $r \geq 0$ and $m \geq 1$, not depending on n, we get*

(a)

$$\mathbb{E}\left(\inf_{i=1,\ldots,n+r} X_i\right)^m = \frac{a_m}{n^m} + \mathcal{O}\Big(\frac{1}{n^{m+1}}\Big),\ a_m > 0;$$

(b)

$$\mathbb{E}\left(1 - \sup_{i=1,\ldots,n+r} X_i\right)^m = \frac{b_m}{n^m} + \mathcal{O}\Big(\frac{1}{n^{m+1}}\Big),\ b_m > 0.$$

(c)

$$\mathbb{E}\left(X_2^* - X_1^*\right) = \frac{d_1}{n} + \mathcal{O}\Big(\frac{1}{n^2}\Big),\ d_1 > 0,\ \textit{and}\ \mathbb{E}\left(X_2^* - X_1^*\right)^m = \mathcal{O}\Big(\frac{1}{n^m}\Big),$$

(d)

$$\mathbb{E}\left(X_n^* - X_{n-1}^*\right) = \frac{e_1}{n} + \mathcal{O}\Big(\frac{1}{n^2}\Big),\ e_1 > 0,\ \textit{and}\ \mathbb{E}\left(X_n^* - X_{n-1}^*\right)^m = \mathcal{O}\Big(\frac{1}{n^m}\Big).$$

Proof. (a) We may write

$$\mathbb{E}\big(\inf_{i=1,\dots,n+r} X_i\big)^m = (n+r)\int_0^1 x^m f(x)\big(1-F(x)\big)^{n+r-1}\,\mathrm{d}x$$
$$= m\int_0^1 x^{m-1}(1-F(x))^{n+r}\,\mathrm{d}x.$$

For $m=1$, we get by condition A3.1-(i) that

$$\frac{1}{(n+r+1)\,\|f\|_\infty} \le \int_0^1 \big(1-F(x)\big)^{n+r}\,\mathrm{d}x \le \frac{1}{c_0(n+r+1)}$$

which implies in turn that there exists $a_1>0$ such that

$$\int_0^1 \big(1-F(x)\big)^{n+r}\,\mathrm{d}x = \frac{a_1}{n} + \mathcal{O}\big(\frac{1}{n^2}\big).$$

For $m\ge 2$, the result follows by induction.
(b) The proof is similar starting from

$$\mathbb{E}\big(1-\sup_{i=1,\dots,n+r} X_i\big)^m$$
$$= (n+r)\int_0^1 (1-x)^m f(x)F^{n+r-1}(x)\,\mathrm{d}x = m\int_0^1 (1-x)^{m-1}F^{n+r}(x)\,\mathrm{d}x.$$

(c) From the joint density of (X_1^*, X_2^*) (see (A.16) or eg David and Nagaraja, 2003, p. 12), we have

$$\mathbb{E}\,(X_2^*-X_1^*)^m = \int_0^1\int_x^1 (y-x)^m n(n-1)f(x)f(y)(1-F(y))^{n-2}\,\mathrm{d}y\,\mathrm{d}x$$
$$= nm\int_0^1\int_0^y (y-x)^{m-1}f(x)(1-F(y))^{n-1}\,\mathrm{d}x\,\mathrm{d}y.$$

For $m>1$, we may bound the last term by

$$\|f\|_\infty\, n\int_0^1 y^m(1-F(y))^{n-1}\,\mathrm{d}y = \|f\|_\infty \frac{n}{m+1}\int_0^1 \mathbb{P}(\inf_{i=1,\dots,n-1} X_i > t^{\frac{1}{m+1}})\,\mathrm{d}t$$
$$= \|f\|_\infty \frac{n}{m+1}\mathbb{E}\big((\inf_{i=1,\dots,n-1} X_i)^{m+1}\big)$$
$$= \mathcal{O}\Big(\frac{1}{n^m}\Big) \text{ by (a).}$$

Also, for $m=1$, we have

$$\mathbb{E}\,(X_2^*-X_1^*) = n\int_0^1 F(y)(1-F(y))^{n-1}\,\mathrm{d}y = n\int_0^1 \frac{t(1-t)^{n-1}}{f(F^{-1}(t))}\,\mathrm{d}t$$

and (c) may be deduced from

$$\frac{n}{\|f\|_\infty}\int_0^1 t(1-t)^{n-1}\,\mathrm{d}t \le n\int_0^1 \frac{t(1-t)^{n-1}}{f(F^{-1}(t))}\,\mathrm{d}t \le \frac{n}{c_0}\int_0^1 t(1-t)^{n-1}\,\mathrm{d}t$$

where $n\int_0^1 t(1-t)^{n-1}\,\mathrm{d}t = \frac{1}{n+1}$.
(d) The last assertions follow from the joint density of (X_{n-1}^*, X_n^*), see (A.16), leading to

$$\mathbb{E}\,(X_n^*-X_{n-1}^*)^m = \int_0^1\int_0^y (y-x)^m n(n-1)f(x)f(y)F^{n-2}(x)\,\mathrm{d}x\,\mathrm{d}y$$
$$= nm\int_0^1\int_x^1 (y-x)^{m-1}F^{n-1}(x)f(y)\,\mathrm{d}y\,\mathrm{d}x.$$

□ □

Now, we are in position to derive the following result with asymptotic equivalents for the integrated bias of the estimators $G_n^{(j,p)}$, $j = 1, 2$.

Proposition 3.3. *Suppose that the condition A3.1-(i) holds, then for all* $p \in [0,1]$ *and constants* a_1, b_1, d_1 *and* e_1 *defined in Lemma 3.2, we have:*
(a)

$$\mathbb{E}\int_0^1 \big(G_n^{(1,p)}(t) - F(t)\big)\,\mathrm{d}t = \frac{p\mathbb{E}\,(X_1^*) + (1-p)\mathbb{E}\,(X_n^*) - p}{2n} = \frac{1-2p}{2n} + \frac{pa_1 - (1-p)b_1}{2n^2} + \mathcal{O}(\frac{1}{n^3});$$

(b)

$$\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(2,p)}(t) - F(t)\big)\,\mathrm{d}t = \frac{(1-p)^2\mathbb{E}\,(X_2^* - X_1^*) + (1-2p)\mathbb{E}\,(X_n^* - X_1^*)}{2n} - \frac{\mathbb{E}\,(X_n^* - X_{n-1}^*)p^2}{2n} = \frac{1-2p}{2n} - \frac{(a_1+b_1)(1-2p)}{2n^2} + \frac{d_1(1-p)^2 - e_1p^2}{2n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big).$$

Proof. Since F_n is unbiased for estimating F we have, for $j = 1, 2$,

$$\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{j,p}(t) - F(t)\big)\,\mathrm{d}t = \mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{j,p}(t) - F_n(t)\big)\,\mathrm{d}t$$

and results are straightforward from Lemma 3.1 ($m = 1$) and Lemma 3.2. □ □

One may remark that these estimators are asymptotically unbiased. For $p = \frac{1}{2}$, the bias is minimal and of order n^{-2} while other values of p give only an integrated bias of order n^{-1}.

3.2. **Mean integrated squared error of polygonal estimators.** Concerning the mean integrated squared error (MISE), we consider its decomposition with respect to F_n:

$$\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(j,p)}(t) - F(t)\big)^2\,\mathrm{d}t = \mathbb{E}\int_{-\infty}^{\infty} \big(F_n(t) - F(t)\big)^2\,\mathrm{d}t + \mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(j,p)}(t) - F_n(t)\big)^2\,\mathrm{d}t + 2\,\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(j,p)}(t) - F_n(t)\big)\big(F_n(t) - F(t)\big)\,\mathrm{d}t,\ j = 1, 2,\ p \in [0,1]. \quad (6)$$

Note that these integrals exist as $X_1, \ldots, X_n$ are supposed to be compactly supported on $[0,1]$. By this way, the range of integration can be taken as $[0,1]$ for $j = 1$ and $\Big[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\Big]$ when $j = 2$. The first term represents the error of reference:

$$\mathbb{E}\int_{-\infty}^{\infty} \big(F_n(t) - F(t)\big)^2\,\mathrm{d}t = \frac{\int_0^1 F(t)\big(1 - F(t)\big)\,\mathrm{d}t}{n}. \quad (7)$$

It remains to study the values of p for which the sum of the two last terms in (6) are globally negative. For such values, our families of polygonal estimators should be more efficient than the empirical distribution estimator.

First, $\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(j,p)}(t) - F_n(t)\big)^2\,\mathrm{d}t$, $j = 1, 2$ can be directly deduced from Lemma 3.1 with $m = 2$ and Lemma 3.2.

Proposition 3.4. *Under the condition A3.1-(i), we have for all* $p \in [0,1]$
(a)

$$\mathbb{E}\int_0^1 \big(G_n^{(1,p)}(t) - F_n(t)\big)^2\,\mathrm{d}t = \frac{(1-2p)\big(p\mathbb{E}\,(X_1^*) + (1-p)\mathbb{E}\,(X_n^*)\big) + p^2}{3n^2} = \frac{1-3(1-p)+3(1-p)^2}{3n^2} + \mathcal{O}\big(\frac{1}{n^3}\big);$$

(b)

$$\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(2,p)}(t) - F_n(t)\big)^2\,\mathrm{d}t = \frac{p^3\mathbb{E}\,(X_n^* - X_{n-1}) + \big((1-p)^3+p^3\big)}{3n^2} + \frac{(1-p)^3\mathbb{E}\,(X_2^* - X_1^*) + \mathbb{E}\,(X_n^* - X_1^*)\big((1-p)^3+p^3\big)}{3n^2} = \frac{1-3(1-p)+3(1-p)^2}{3n^2} + \mathcal{O}\big(\frac{1}{n^3}\big).$$

We may remark that, again, the case $p = \frac{1}{2}$ appears as simpler since the calculus of $\mathbb{E}\int_0^1 \big(G_n^{(1,\frac{1}{2})}(t) - F_n(t)\big)^2\,\mathrm{d}t$ reduces to $\frac{1}{12n^2}$. The most difficult task is the study of the double product, we obtain the following result which is proved in the Appendix A.

Proposition 3.5. *Under Assumption 3.1, we get for* $j = 1, 2$ *and* $p \in [0,1]$*:*

$$2\,\mathbb{E}\int_0^1 \big(G_n^{(j,p)}(t) - F_n(t)\big)\big(F_n(t) - F(t)\big)\,\mathrm{d}t = -\frac{1}{3n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big).$$

Now, collecting results in (6)-(7) and Proposition 3.4 and 3.5, we are in position to state the main result of this paper.

Theorem 3.6. *Under Assumption 3.1, we get for* $j = 1, 2$ *and all* $p \in [0,1]$*:*

$$\mathbb{E}\int_{-\infty}^{\infty} \big(G_n^{(j,p)}(t) - F(t)\big)^2\,\mathrm{d}t = \frac{1}{n}\int_0^1 F(t)\big(1-F(t)\big)\,\mathrm{d}t - \frac{p(1-p)}{n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big).$$

First, we may conclude that for all $p \in [0,1]$, estimators $G_n^{(1,p)}$ and $G_n^{(2,p)}$ are asymptotically equivalent. Indeed, $G_n^{(2,p)}$ is only a slight modification of $G_n^{(1,p)}$ at its extremities, so this result seems natural. Next, for all $p \in (0,1)$, the families $G_n^{(j,p)}$, $j = 1, 2$ appear as more efficient than the empirical distribution estimator F_n. The choices $p = 0$ or $p = 1$ turn out to be more problematic since the term $\frac{p(1-p)}{n^2}$ vanishes in these cases. If these estimators improve F_n, the gain can only occur for the third order what seems to be of less interest. Finally, this latter term is maximal for $p = \frac{1}{2}$ with the value $\frac{1}{4n^2}$. In conclusion, among all the family of polygonal estimators, one should prefer to work with $G_n^{(1,\frac{1}{2})}$ or $G_n^{(2,\frac{1}{2})}$ (in the case of unknown support $[a,b]$) because these estimators have both the smaller asymptotic bias by Proposition 3.3 and the better efficiency relative to the classical distribution estimator F_n.

4. Simulation

4.1. **Framework.** In this section, we look at the small sample behaviour of the polygonal estimators, $G_n^{(j,p)}$, $j = 1, 2$ with a focus on the values $p = \frac{1}{2}$ and $p = 0$ or 1 (represented on Figure 1). To this aim, we consider the set of 15 Gaussian mixtures defined in Marron and Wand (1992), denoted by MW1-MW15 in the sequel, and also, the 16th density introduced in Janssen et al. (1995), say MW16. The parameters of these normal mixtures are recalled in the Appendix B. These

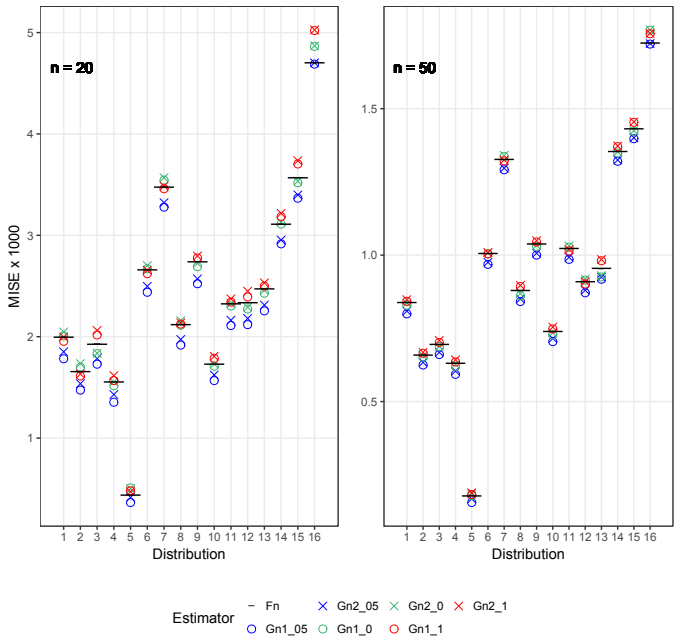

FIGURE 2. MISE for F_n, $G_n^{(j,0)}$, $G_n^{(j,\frac{1}{2})}$, $G_n^{(j,1)}$, $j = 1, 2$, for $n = 20$ (left) and $n = 50$ (right)

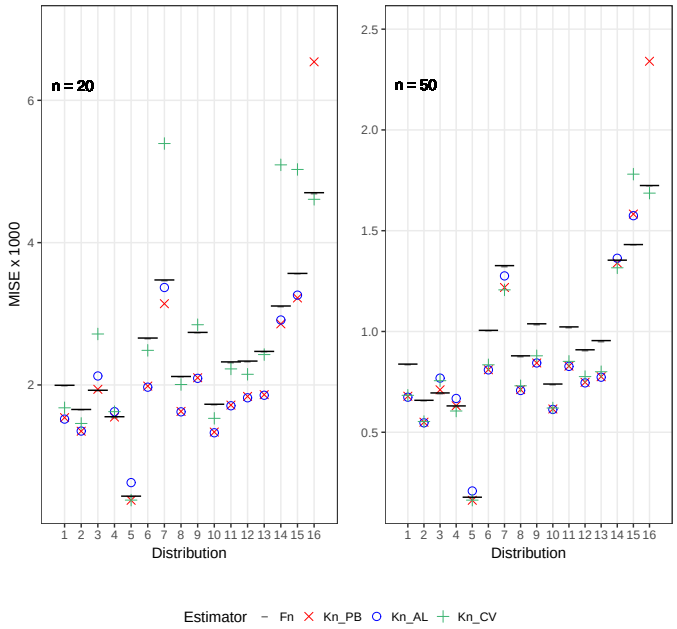

FIGURE 3. MISE for F_n and kernel distribution estimators with three bandwidth choices : AL, PB and CV for $n = 20$ (left) and $n = 50$ (right)

distributions are easy to implement and describe a broad class of potential problems that may occur in nonparametric estimation (skewness, multimodality, and heavy kurtosis). As their parameters were chosen such that $\min_\ell(\mu_\ell - 3\sigma_\ell) = -3$ and $\max_\ell(\mu_\ell + 3\sigma_\ell) = 3$, the estimators $G_n^{(1,p)}$ are computed with the values $a = -3$ and $b = 3$. For each of these distributions, $N = 500$ samples of sizes $n = 20$, 50 and 100 are generated with the R software (R Core Team, 2017). Next, a Monte Carlo approximation, based on 14000 trials over $[-7, 7]$, is operated for each sample of size n to estimate $\int_{-7}^{7}(G_n^{(j,p)} - F(t))^2\,\mathrm{d}t$. The final approximation of the MISE, $\widehat{\mathrm{MISE}}$, is obtained by averaging the results over the $N = 500$ replicates.

4.2. **Results for the polygonal estimators.** Approximations of the MISE for F_n and estimators $G_n^{(j,p)}$, $j \in \{1, 2\}$, $p \in \{0, \frac{1}{2}, 1\}$ are reported in Figure 2. Here, only the cases $n = 20$ and $n = 50$ are represented since for $n = 100$ results are similar but with almost indistinguishable differences between the estimators. First of all, we remark that the obtained results are in accordance with the theoretical results of Theorem 3.6: the estimators $G_n^{(j,\frac{1}{2})}$ have a smaller MISE than F_n for all simulated distributions. In addition, the estimator $G_n^{(1,\frac{1}{2})}$ calculated with the choice $[a, b] = [-3, 3]$ gives slight better results than $G_n^{(2,\frac{1}{2})}$. So even if $[-3, 3]$ is not the true support of our simulated distributions, the knowledge of this prediction interval is sufficient to improve the estimation. Concerning $G_n^{(j,0)}$ and $G_n^{(j,1)}$, results are clearly inferior to those of $G_n^{(j,\frac{1}{2})}$. In accordance with our theoretical results, the use of these estimators should not be recommended (even if $G_n^{(j,0)}$ seems to achieve globally better results than $G_n^{(j,1)}$).

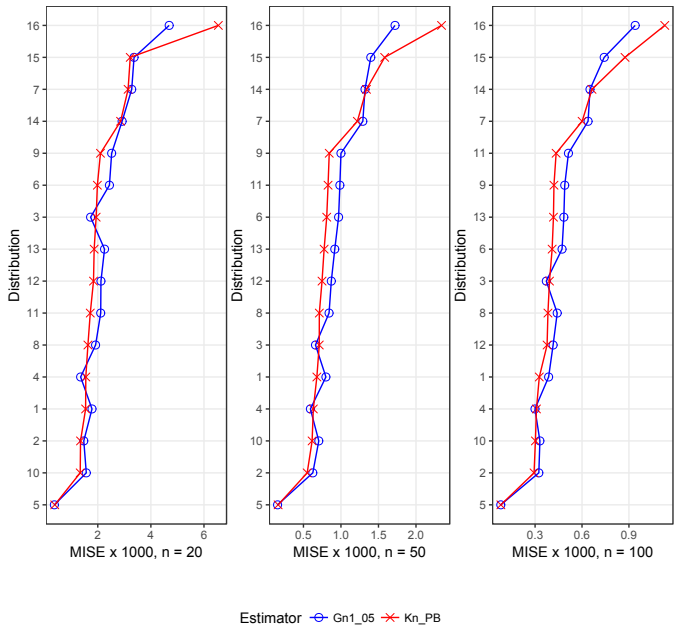

FIGURE 4. Compared MISE of $G_n^{(1,\frac{1}{2})}$ and $K_{n,\mathrm{PB}}$ for $n = 20$, 50 and 100.

4.3. **Comparison with kernel distribution estimators.** To complete our simulation study, we also compare our estimators with the nonparametric kernel distribution estimator defined by

$$K_n(t) = \frac{1}{nh_n}\sum_{i=1}^{n} L\big(\frac{t-X_i}{h_n}\big),\ t\in\mathbb{R}$$

where h_n is the bandwidth and $L(t)=\int_{-\infty}^{t} K(x)\,\mathrm{d}x$. Here K is the usual kernel used in density estimation, chosen as a known continuous density on $\mathbb{R}$, symmetric about 0. Theoretical properties of this estimator are well known, one may refer e.g. to Swanepoel and Van Graan (2005) or Servien (2009) with a rich literature review. The weighted MISE of K_n, $\mathbb{E}\int_{-\infty}^{\infty}\big(K_n(t)-F(t)\big)^2 f(t)\,\mathrm{d}t$ is established in Swanepoel (1988) where optimal choice of the kernel K is discussed. The unweighted case is derived in Jones (1990) when F has two continuous derivatives f and f':

$$\mathbb{E}\int_{-\infty}^{\infty}\big(K_n(t)-F(t)\big)^2\,\mathrm{d}t = \frac{\int_{-\infty}^{\infty}F(t)(1-F(t))\,\mathrm{d}t}{n} - \frac{2h_n}{n}\int_{-\infty}^{\infty}tK(t)L(t)\,\mathrm{d}t + \frac{h_n^4}{4}\Big(\int_{-\infty}^{\infty}t^2K(t)\,\mathrm{d}t\Big)^2\int_{-\infty}^{\infty}\big(f'(t)\big)^2\,\mathrm{d}t + o\big(h_n^4\big)+o\big(\frac{h_n}{n}\big).$$

Actually, if one considers only the Lipschitz condition A3.1-(ii) given on f, straightforward calculation with Taylor series yields that the latter result is weakened in

$$\mathbb{E}\int_{0}^{1}\big(K_n(t)-F(t)\big)^2\,\mathrm{d}t = \frac{\int_{0}^{1}F(t)(1-F(t))\,\mathrm{d}t}{n} - \frac{2h_n}{n}\int_{-\infty}^{\infty}tK(t)L(t)\,\mathrm{d}t + \mathcal{O}\big(h_n^4\big)+o\big(\frac{h_n}{n}\big). \quad (8)$$

Comparing (8) with results of Theorem 3.6, it appears that the expressions are similar with the presence of the MISE of F_n in both of ones. Also, for h_n of order $n^{-\frac{1}{3}}$, the improvement toward F_n stands only on the 2nd order effect: about $n^{-4/3}$ for K_n; but only n^{-2} for the estimators $G_n^{(j,p)}$ when $p\in(0,1)$. Indeed, K_n should be asymptotically better but, the following numerical results show that exceptions hold for some densities. As usual, choosing the bandwidth h_n is a critical task to avoid over or under-smoothing of the data. Several procedures have been proposed in the literature for kernel distribution estimation, we may refer among others to Sarda (1993) for a leave-one-out cross-validation method, Altman and Léger (1995) or Polansky and Baker (2000) for a plug-in bandwidth choice and Bowman et al. (1998) for a modified cross-validation method. First, we compare F_n to these estimators respectively called $K_{n,\mathrm{AL}}$, $K_{n,\mathrm{PB}}$ and $K_{n,\mathrm{CV}}$ in the following. To this end, we use in R the `kerdiest` package of Quintela-del Río and Estévez-Pérez (2012). The selection of K appears as less crucial and we take the normal kernel in our simulations.

It appears in Figure 3 that for $n=20$ and $n=50$, the plug-in bandwidth choices of Altman and Léger (1995) and Polansky and Baker (2000) give very similar results except than for MW16 where $K_{n,\mathrm{AL}}$ performs quite badly (with an approximate MISE beyond the frame). Also, the cross-validation proposed by Bowman et al. (1998) seems to be less recommandable for these sample sizes. We may observed that for most of the distributions, $K_{n,\mathrm{AL}}$ and $K_{n,\mathrm{PB}}$ have a smaller MISE than F_n. Exceptions hold for $n=20$ with the two distributions MW3 and MW16 and, MW15 in addition for $n=50$. Again, the case $n=100$ is not represented but results are similar for $K_{n,\mathrm{AL}}$ and $K_{n,\mathrm{PB}}$ (with the same exceptions). Also, we observe that $K_{n,\mathrm{CV}}$ gives much more satisfactory results.

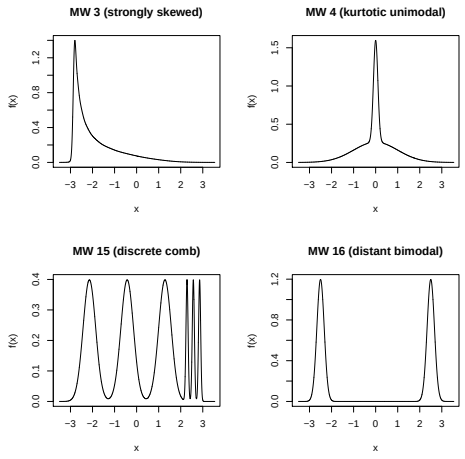

FIGURE 5. Densities of 4 selected MW distributions.

4.4. **Comparison of $G_n^{(1,\frac{1}{2})}$ and $K_{n,\mathbf{PB}}$.** In this part, we focus on the Polansky and Baker (2000) bandwidth choice (with multistage plug-in) because it is fastest in terms of computation time and also, it gives the most homogeneous results for the tested distributions. Indeed, since polygonal estimators do not depend on any bandwidth, it seems fairer to select one particular bandwidth choice rather than to adapt the method according to the density or the sample size. In Figure 4, the MISE of this kernel estimator (ordered by increasing value) is compared to the best polygonal estimator $G_n^{(1,\frac{1}{2})}$. For $n = 20$, 50 and 100, the obtained results are quite close and $K_{n,\mathrm{PB}}$ performs better for almost all distributions. This is not surprising from the result derived in (8). The sensitivity toward a bad bandwidth choice appears for the distributions MW3, MW4, MW15 and MW16 where the polygon estimator achieves better results. The respective densities are drawn in Figure 5 and the obtained estimations for $G_n^{(1,\frac{1}{2})}$ and $K_{n,\mathrm{PB}}$ are given in Figure 6 for $n = 50$. It appears that the estimates are quite close but in these special cases, the kernel estimator misses the curvatures. Finally, recall that $G_n^{(1,\frac{1}{2})}$ is constructed with the help of the prediction interval [-3,3] and the full nonparametric estimator $G_n^{(2,\frac{1}{2})}$ differs from $G_n^{(1,\frac{1}{2})}$ only at both ends. Lower tail of estimated distributions are zoomed in Figure 7 to compare these two estimators.

5. Discussion

We have studied two general families $G_n^{(1,p)}$ and $G_n^{(2,p)}$ of smoothed polygonal distribution estimators and have derived their properties as well as exact expansions for the MISE at the second order for compactly supported distributions. These estimators present several advantages: they can be derived directly from the empirical distribution function F_n, they do not depend on any smoothing parameter and may be hand drawn. Because of their proximity to F_n, they also inherit its convergence properties. Their study fills a gap since these estimators are quite naturally used by practitioners but their theoretical properties had not yet been studied in depth. Examination of the second-order effect in the MISE shows that $G_n^{(1,p)}$ and $G_n^{(2,p)}$ are asymptotically equivalent. Also, the MISE of F_n is improved for all p chosen in $(0, 1)$, and its minimal value is reached for p equals to $\frac{1}{2}$. On the other hand, our results, for $p = 0$ or 1, also show that joining the ends of F_n does not necessarily

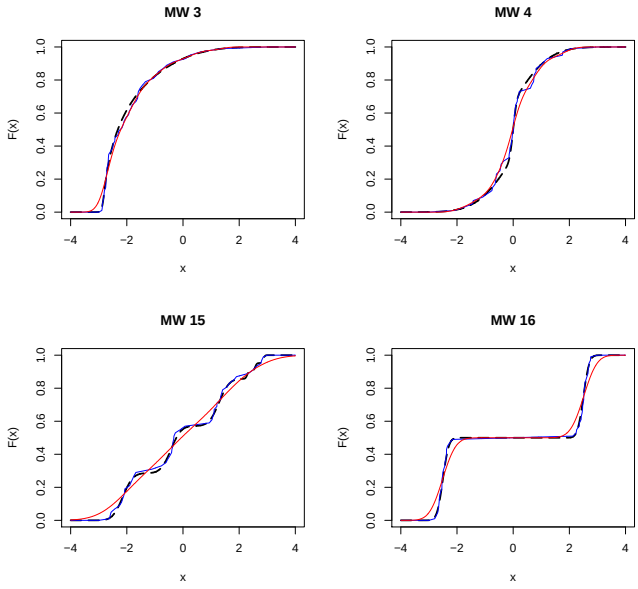

FIGURE 6. Estimations of F (dashed) with $G_n^{(1,\frac{1}{2})}$ (blue) and $K_{n,\mathrm{PB}}$ (red) for the four selected MW distributions and $n = 50$

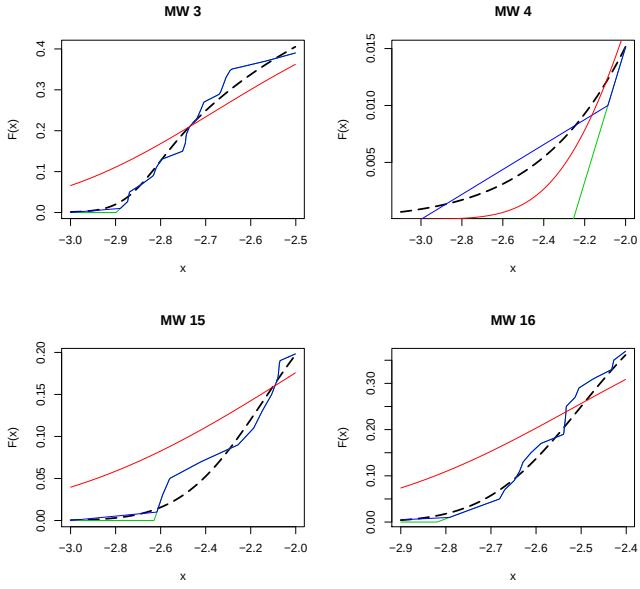

FIGURE 7. Same estimations of F (dashed) zoomed on small values of x with $G_n^{(1,\frac{1}{2})}$ (blue), $G_n^{(2,\frac{1}{2})}$ (green), $K_{n,\mathrm{PB}}$ (red) and $n = 50$

improve the estimation and indeed may be worsen it. Our simulations on Gaussian mixtures support these conclusions and also show that $G_n^{(1,\frac{1}{2})}$ may achieve better results than $G_n^{(2,\frac{1}{2})}$ even for distributions with infinite support.

Various extensions of these results may be envisaged. A first one is to relax the assumption of bounded support and then, to consider a weighted mean integrated squared error to ensure the existence of the integrals. Remark that the family of estimators $G_n^{(2,p)}$ is naturally adapted to this framework. This approach involves not only technical difficulties for the study of the integrals but also for establishing the analog of Lemma 3.2 where the asymptotic behaviour of extremal order statistics has to be studied, see e.g. David and Nagaraja (2003, chapter 4) and references therein.

Another possible extension is to consider the non-iid case where $X_1, \ldots, X_n$ are only identically distributed random variables. Examination of our proofs shows that the results mainly depend on the bounds obtained for $\mathbb{E}(X_1^*)$, $\mathbb{E}(X_2^* - X_1^*)$, $\mathbb{E}(X_n^* - X_{n-1}^*)$ and $\mathbb{E}(X_n^*)$, and, on Proposition A.1 (derived in the Appendix A). We infer that some of our results can be straightforward generalized to the case of exchangeable variables, ie, with a joint cumulative distribution function supposed to be symmetric in its arguments. Concerning the extension of Proposition A.1, a first step would be the study of $f_{(X_k^*, X_{k+1}^*)}$. To this end, results obtained by Maurer and Margolin (1976); Caraux and Gascuel (1992); Rychlik (1993, 1994) and Kaluszka and Okolewski (2001) should be interesting.

In the framework of stochastic processes, a last possibility may be envisaged concerning the smoothed Poisson process where, proceeding as in (1)-(2), one may link the process between successive arrival times. An alternative should be also the spline smoothing with derivatives. This allows to consider this smoothed process as a functional process with values in $C([0, +\infty))$ and to consider limit theorems and exponential inequalities in this space. We refer to Bosq (2017) for further developments in this direction.

The authors gratefully thank the referees and editors for their constructive and insightful comments on the manuscript.

References

N. Altman and C. Léger. Bandwidth selection for kernel distribution function estimation. *J. Statist. Plann. Inference*, 46(2):195–214, 1995.

A. Azzalini. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68(1):326–328, 1981.

G. J. Babu, A. J. Canty, and Y. P. Chaubey. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Statist. Plann. Inference*, 105(2):377–392, 2002.

P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999.

D. Bosq. Predicting smoothed Poisson process and regularity for density estimation in the context of an exponential rate. In preparation, 2017.

A. Bowman, P. Hall, and T. Prvan. Bandwidth selection for the smoothing of distribution functions. *Biometrika*, 85(4):799–808, 1998.

G. Caraux and O. Gascuel. Bounds on distribution functions of order statistics for dependent variates. *Statist. Probab. Lett.*, 14(2):103–105, 1992.

M.-Y. Cheng and L. Peng. Regression modeling for nonparametric estimation of distribution and quantile functions. *Statist. Sinica*, 12(4):1043–1060, 2002.

H. A. David and H. N. Nagaraja. *Order statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

M. L. Huang and P. H. Brill. A distribution estimation method based on level crossings. *J. Statist. Plann. Inference*, 124(1):45–62, 2004.

P. Janssen, J. S. Marron, N. Veraverbeke, and W. Sarle. Scale measures for bandwidth selection. *J. Nonparametr. Statist.*, 5(4):359–380, 1995.

M. C. Jones. The performance of kernel density functions in kernel distribution function estimation. *Statist. Probab. Lett.*, 9(2):129–132, 1990.

M. Kaluszka and A. Okolewski. An extension of the Erdős-Neveu-Rényi theorem with applications to order statistics. *Statist. Probab. Lett.*, 55(2):181–186, 2001.

A. Leblanc. On estimating distribution functions using Bernstein polynomials. *Ann. Inst. Statist. Math.*, 64(5):919–943, 2012.

M. Lejeune and P. Sarda. Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.*, 14(4):457–471, 1992.

J. S. Marron and M. P. Wand. Exact mean integrated squared error. *Ann. Statist.*, 20(2):712–736, 1992.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.

W. Maurer and B. H. Margolin. The multivariate inclusion-exclusion formula and order statistics from dependent variates. *Ann. Statist.*, 4(6):1190–1199, 1976.

A. M. Polansky and E. R. Baker. Multistage plug-in bandwidth selection for kernel distribution function estimates. *J. Statist. Comput. Simulation*, 65(1):63–80, 2000.

A. Quintela-del Río and G. Estévez-Pérez. Nonparametric kernel distribution function estimation with kerdiest: an R Package for bandwidth choice and applications. *Journal of Statistical Software*, 50(8):1–21, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

R. R. Read. The asymptotic inadmissibility of the sample distribution function. *Ann. Math. Statist.*, 43:89–95, 1972.

T. Rychlik. Sharp bounds on L-estimates and their expectations for dependent samples. *Comm. Statist. Theory Methods*, 22(4):1053–1068, 1993.

T. Rychlik. Distributions and expectations of order statistics for possibly dependent random variables. *J. Multivariate Anal.*, 48(1):31–42, 1994.

P. Sarda. Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference*, 35(1):65–75, 1993.

D. W. Scott. Frequency polygons: theory and application. *J. Amer. Statist. Assoc.*, 80(390):348–354, 1985.

R. Servien. Estimation de la fonction de répartition : revue bibliographique. *J. SFdS*, 150(2):84–104, 2009.

J. S. Simonoff. *Smoothing methods in statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

J. Swanepoel. Mean integrated squared error properties and optimal kernels when estimating a distribution function. *Comm. Statist. Theory Methods*, 17(11):3785–3799, 1988.

J. Swanepoel and F. Van Graan. A new kernel distribution function estimator based on a non-parametric transformation of the data. *Scand. J. Statist.*, 32(4): 551–562, 2005.

Appendix Appendix A. Proof of Proposition 3.5

To prove Proposition 3.5, we start from equation (6) and consider, for $j = 1, 2$, its decomposition in two terms:

$$2\,\mathbb{E}\int_0^1 \big(G_n^{(j,p)}(t) - F_n(t)\big)\big(F_n(t) - F(t)\big)\,\mathrm{d}t = I_1^{(j,p)} + I_2^{(j,p)}$$

with

$$I_1^{(j,p)} = 2\,\mathbb{E}\int_{-\infty}^{+\infty} \big(G_n^{(j,p)}(t) - F_n(t)\big)F_n(t)\,\mathrm{d}t \tag{A.9}$$

and

$$I_2^{(j,p)} = 2\,\mathbb{E}\int_{-\infty}^{+\infty} \big(F_n(t) - G_n^{(j,p)}(t)\big)F(t)\,\mathrm{d}t. \tag{A.10}$$

Note that for $j = 1$, the range of integration is reduced to the interval $[0, 1]$.

A.1. **Study of $I_1^{(1,p)}$.** From (3), we have

$$2\,\big(G_n^{(1,p)}(t) - F_n(t)\big)F_n(t) = \sum_{k=1}^{n-1} \frac{2kt - 2pkX_{k+1}^* - 2(1-p)kX_k^*}{n^2(X_{k+1}^* - X_k^*)}\,\mathbb{I}_{[X_k^*, X_{k+1}^*)}(t) - \frac{2p(1-t)}{n(1-X_n^*)}\,\mathbb{I}_{[X_n^*,1]}(t).$$

The calculation of the integral leads to

$$I_1^{(1,p)} = -\mathbb{E}\Big(\sum_{k=1}^{n-1} \frac{(2p-1)k(X_{k+1}^* - X_k^*)}{n^2}\Big) - \frac{p(1 - \mathbb{E}\,(X_n^*))}{n}$$

so one may conclude that

$$I_1^{(1,p)} = \frac{-p - (1-2p)\mathbb{E}\,(X_1) + (1-p)\mathbb{E}\,(X_n^*)}{n}$$

and Lemma 3.2-(b) gives in turn

$$I_1^{(1,p)} = \frac{(1-2p)(1 - \mathbb{E}\,(X_1))}{n} - \frac{b_1(1-p)}{n^2} + \mathcal{O}\big(\frac{1}{n^3}\big). \tag{A.11}$$

A.2. **Study of $I_2^{(1,p)}$.** It is the most difficult term. Again from (3), we may derive:

$$2\,\big(F_n(t) - G_n^{(1,p)}(t)\big)F(t) = -\frac{2(1-p)tF(t)}{nX_1^*}\,\mathbb{I}_{[0,X_1^*)}(t) + \frac{2p(F(t) - tF(t))}{n(1-X_n^*)}\,\mathbb{I}_{[X_n^*,1]}(t) - \frac{2}{n}\sum_{k=1}^{n-1} \frac{tF(t) - \big(pX_{k+1}^* + (1-p)X_k^*\big)F(t)}{X_{k+1}^* - X_k^*}\,\mathbb{I}_{[X_k^*, X_{k+1}^*)}(t).$$

Denoting by $K_0(t)$ the primitive of $F(t)$ and $K_1(t)$ that of $tF(t)$, we get

$$\begin{aligned} &2\int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t \\ &\quad = -\frac{2(1-p)\big(K_1(X_1^*) - K_1(0)\big)}{nX_1^*} + \frac{2p\big(K_0(1) - K_0(X_n^*) - K_1(1) + K_1(X_n^*)\big)}{n(1-X_n^*)} \\ &- \frac{2}{n}\sum_{k=1}^{n-1} \frac{\big(K_1(X_{k+1}^*) - K_1(X_k^*)\big) - \big(pX_{k+1}^* + (1-p)X_k^*\big)\big(K_0(X_{k+1}^*) - K_0(X_k^*)\big)}{X_{k+1}^* - X_k^*}. \end{aligned} \tag{A.12}$$

Setting $X_0^* = 0$ and $X_{n+1}^* = 1$, we have also for all $k = 0, \ldots, n$:

$$K_0(X_{k+1}^*) - K_0(X_k^*) = (X_{k+1}^* - X_k^*)F(X_k^*) + \int_{X_k^*}^{X_{k+1}^*} f(t)(X_{k+1}^* - t)\,\mathrm{d}t \tag{A.13}$$

and, after integration by parts,

$$K_1(X^*_{k+1}) - K_1(X^*_k) = \frac{1}{2}\big((X^*_{k+1})^2 - (X^*_k)^2\big)F(X^*_k) + \frac{1}{2}\int_{X^*_k}^{X^*_{k+1}} f(t)\big((X^*_{k+1})^2 - t^2\big)\,\mathrm{d}t. \tag{A.14}$$

We report (A.13) and (A.14) in (A.12) to obtain with $M^*_k = X^*_k + \theta_k(X^*_{k+1} - X^*_k)$, $0 < \theta_k < 1$ for $k = 0, \ldots, n$:

$$\begin{aligned} 2\int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t &= \frac{1}{n}\sum_{k=1}^{n-1}(2p-1)(X^*_{k+1} - X^*_k)F(X^*_k) \\ &\quad + \frac{-2(1-p)(X^*_1)^2 f(M^*_0) + 3pF(X^*_n)(1-X^*_n) + pf(M^*_n)(1-X^*_n)^2}{3n} \\ &\quad + \frac{1}{n}\sum_{k=1}^{n-1}\int_{X^*_k}^{X^*_{k+1}} f(t)\Big(\frac{X^*_{k+1}-t}{X^*_{k+1}-X^*_k}\Big)\big((X^*_{k+1}-t) - 2(1-p)(X^*_{k+1}-X^*_k)\big)\,\mathrm{d}t. \end{aligned}$$

giving in turn

$$\begin{aligned} &2\int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t \\ &\quad = \frac{-2(1-p)(X^*_1)^2 f(M^*_0) + 3pF(X^*_n)(1-X^*_n) + pf(M^*_n)(1-X^*_n)^2}{3n} \\ &\qquad + \frac{1}{n}\sum_{k=1}^{n-1}(2p-1)(X^*_{k+1} - X^*_k)F(X^*_k) + \frac{3p-2}{3}f(M^*_k)(X^*_{k+1} - X^*_k)^2. \end{aligned}$$

Finally, the condition A3.1-(ii) leads to:

$$\begin{aligned} I_2^{(1,p)} &= 2\int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t = \frac{-2(1-p)(X^*_1)^2\big(f(0) + X^*_1 R_{1,0}\big)}{3n} \\ &\quad + \frac{p(1-X^*_n)\big(3F(X^*_n) + (1-X^*_n)(f(X^*_n) + (1-X^*_n)R_{1,n})\big)}{3n} \\ &+ \frac{(2p-1)}{n}\sum_{k=1}^{n-1}(X^*_{k+1} - X^*_k)F(X^*_k) + \frac{(3p-2)}{3n}\sum_{k=1}^{n-1}(X^*_{k+1} - X^*_k)^2\big(f(X^*_k) + (X^*_{k+1} - X^*_k)R_{1,k}\big) \end{aligned} \tag{A.15}$$

with $|R_{1,k}| \le c_1\theta_k < c_1$, $k = 0, \ldots, n$.

Next, the following proposition will be useful for further calculations. It is obtained with the binomial theorem applied to the joint density of (X^*_k, X^*_{k+1}) (see e.g. David and Nagaraja, 2003, p. 12) given by

$$f_{(X^*_k, X^*_{k+1})}(x,y) = \frac{n!}{(k-1)!(n-k-1)!}F^{k-1}(x)f(x)f(y)(1-F(y))^{n-k-1}\mathbb{I}_{[0,y]}(x)\mathbb{I}_{[0,1]}(y). \tag{A.16}$$

Proposition A.1. *If h is measurable and integrable on $[0,1]^2$, then*

$$\sum_{k=1}^{n-1}\mathbb{E}\left(h(X^*_k, X^*_{k+1})\right) = n(n-1)\int_0^1\int_0^y h(x,y)f(x)f(y)\big(1-F(y)+F(x)\big)^{n-2}\,\mathrm{d}x\,\mathrm{d}y.$$

For expressions depending on k in (A.15), we get from Proposition A.1 and after integration by parts, that

$$\sum_{k=1}^{n-1}\mathbb{E}\left((X_{k+1}^*-X_k^*)F(X_k^*)\right)=\int_0^1 F(x)\,\mathrm{d}x-\frac{1}{n+1}-\int_0^1\frac{nF^{n+1}(x)}{n+1}\,\mathrm{d}x \\ +\int_0^1\frac{(1-F(y))^{n+1}}{n+1}\,\mathrm{d}y,$$

which, with Lemma 3.2, can be written as

$$\sum_{k=1}^{n-1}\mathbb{E}\left((X_{k+1}^*-X_k^*)F(X_k^*)\right)=1-\mathbb{E}\,(X_1)-\frac{1}{n+1}-\frac{n}{n+1}\mathbb{E}\,(1-\sup_{i=1,\dots,n+1}X_i) \\ +\frac{1}{n+1}\mathbb{E}\,(\inf_{i=1,\dots,n+1}X_i)=1-\mathbb{E}\,(X_1)-\frac{1}{n}-\frac{b_1}{n}+\mathcal{O}(\frac{1}{n^2}). \quad \text{(A.17)}$$

Next,

$$\sum_{k=1}^{n-1}\mathbb{E}\left((X_{k+1}^*-X_k^*)^2f(X_k^*)\right)=-\int_0^1 nf^2(x)(1-x)^2F^{n-1}(x)\,\mathrm{d}x \\ +2n\int_0^1\int_0^y f^2(x)(y-x)(1-F(y)+F(x))^{n-1}\,\mathrm{d}x\,\mathrm{d}y$$

and from

$$\|f\|_\infty\int_0^1(1-x)^2f(x)nF^{n-1}(x)\,\mathrm{d}x=2\,\|f\|_\infty\int_0^1(1-x)F^n(x)\,\mathrm{d}x \\ =\|f\|_\infty\int_0^1\mathbb{P}(1-X_n^*\geq\sqrt{t})\,\mathrm{d}t=\mathbb{E}\,((1-X_n^*)^2),$$

we get

$$\sum_{k=1}^{n-1}\mathbb{E}\left((X_{k+1}^*-X_k^*)^2f(X_k^*)\right) \\ =2n\int_0^1\int_0^y f^2(x)(y-x)(1-F(y)+F(x))^{n-1}\,\mathrm{d}x\,\mathrm{d}y+\mathcal{O}(\frac{1}{n^2}).$$

As $f^2(x)=f(x)f(y)+f(x)(f(x)-f(y))$, the double integral is equal to

$$2n\int_0^1\int_0^y f(x)f(y)(y-x)(1-F(y)+F(x))^{n-1}\,\mathrm{d}x\,\mathrm{d}y \\ +2n\int_0^1\int_0^y f(x)(f(x)-f(y))(y-x)(1-F(y)+F(x))^{n-1}\,\mathrm{d}x\,\mathrm{d}y.$$

For the first integral, the following expression is obtained:

$$\frac{2}{n+1}-\frac{2}{n+1}\int_0^1\big(1-F(y)\big)^{n+1}\,\mathrm{d}y-\frac{2}{n+1}\int_0^1F^{n+1}(x)\,\mathrm{d}x=\frac{2}{n}+\mathcal{O}(\frac{1}{n^2})$$

by Lemma 3.2. On the other hand, the second integral can be bounded with the condition A3.1-(ii) to obtain the following new bound:

$$2nc_1\int_0^1\int_0^y f(x)(y-x)^2(1-F(y)+F(x))^{n-1}\,\mathrm{d}x\,\mathrm{d}y \\ =4c_1\int_0^1\int_0^y(y-x)(1-F(y)+F(x))^n\,\mathrm{d}x\,\mathrm{d}y+\mathcal{O}(\frac{1}{n^3}).$$

Using the condition A3.1-(i), one may bound this double integral with

$$c_0^{-1}\int_0^1\int_0^y f(x)(y-x)(1-F(y)+F(x))^n\,\mathrm{d}x\,\mathrm{d}y$$

and an integration by part gives

$$\frac{1}{c_0}\int_0^1\frac{y(1-F(y))^{n+1}}{n+1}\,\mathrm{d}y+\frac{1}{c_0}\int_0^1\int_x^1\frac{(1-F(y)+F(x))^{n+1}}{n+1}\,\mathrm{d}y\,\mathrm{d}x$$

$$\leq\frac{\mathbb{E}\,((\inf\limits_{i=1,\ldots,n+1}X_i)^2)}{2c_0(n+1)}-\frac{1-\mathbb{E}\,(\sup\limits_{i=1,\ldots,n+2}X_i)}{c_0^2(n+2)(n+1)}+\frac{1}{c_0^2(n+1)(n+2)}.$$

Finally, we may conclude with Lemma 3.2-(a)(b) that

$$\sum_{k=1}^{n-1}\mathbb{E}\left((X_{k+1}^*-X_k^*)^2f(X_k^*)\right)=\frac{2}{n}+\mathcal{O}(\frac{1}{n^2}).\tag{A.18}$$

Concerning the last term, we have

$$\sum_{k=1}^{n-1}\mathbb{E}\,(X_{k+1}^*-X_k^*)^3=-3\int_0^1(1-x)^2F^n(x)\,\mathrm{d}x-3\int_0^1y^2(1-F(y))^n\,\mathrm{d}y$$
$$+6\int_0^1\int_0^y(y-x)(1-F(y)+F(x))^n\,\mathrm{d}x\,\mathrm{d}y.$$

We obtain

$$\begin{aligned}\sum_{k=1}^{n-1}&\mathbb{E}\,(X_{k+1}^*-X_k^*)^3\\&=-\mathbb{E}\,((1-X_n^*)^3)-\mathbb{E}\,((X_1^*)^3)+6\int_0^1\int_0^y(y-x)(1-F(y)+F(x))^n\,\mathrm{d}x\,\mathrm{d}y\\&=6\int_0^1\int_0^y(y-x)(1-F(y)+F(x))^n\,\mathrm{d}x\,\mathrm{d}y+\mathcal{O}(\frac{1}{n^3})\end{aligned}$$

and, previous calculations lead to

$$\sum_{k=1}^{n-1}\mathbb{E}\,(X_{k+1}^*-X_k^*)^3=\mathcal{O}(\frac{1}{n^2}).\tag{A.19}$$

Noting that $\mathbb{E}\,((1-X_n^*)^2f(X_n^*))=\mathcal{O}(\frac{1}{n^2})$ and

$$\mathbb{E}\,((1-X_n^*)F(X_n^*))=\frac{n}{n+1}\mathbb{E}\,(1-\sup_{i=1,\ldots,n+1}X_i)=\frac{b_1}{n}+\mathcal{O}(\frac{1}{n^2}),$$

we may plug the results in (A.15), together with (A.17)-(A.19), to obtain that:

$$\begin{aligned}I_2^{(1,p)}&=2\,\mathbb{E}\int_0^1\left(F_n(t)-G_n^{(1,p)}(t)\right)F(t)\,\mathrm{d}t\\&=\frac{(2p-1)(1-\mathbb{E}\,(X_1))}{n}+\frac{b_1(1-p)}{n^2}-\frac{1}{3n^2}+\mathcal{O}(\frac{1}{n^3})\end{aligned}\tag{A.20}$$

A.3. **Conclusion for $G_n^{(1,p)}$.** Collecting the results obtained in (A.9)-(A.11) and (A.20), we obtain that

$$2\,\mathbb{E}\int_0^1\left(G_n^{(1,p)}(t)-F_n(t)\right)\left(F_n(t)-F(t)\right)\mathrm{d}t=-\frac{1}{3n^2}+\mathcal{O}(\frac{1}{n^3}),$$

so Proposition 3.5 is proved for $G_n^{(1,p)}$.

A.4. **Study of $I_1^{(2,p)}$.** Concerning the estimators $G_n^{(2,p)}$, proofs are similar but with more complex terms due to the random integration bounds. Here and in the following, we point out only the main changes. For the term $I_1^{(2,p)}$ defined in (A.9), first we have

$$2\big(G_n^{(2,p)}(t)-F_n(t)\big)F_n(t)=2\big(G_n^{(1,p)}(t)-F_n(t)\big)F_n(t)\big)\,\mathbb{I}_{[X_1^*,X_n^*)}(t) + \frac{t-(1+p)X_n^*+pX_{n-1}^*}{n(X_n^*-X_{n-1}^*)}\,\mathbb{I}_{[X_n^*,(1+p)X_n^*-pX_{n-1}^*)}(t).$$

The calculation of the integrals and their expectation lead to

$$I_1^{(2,p)}=\frac{-(2p-1)(1-\mathbb{E}\,(X_1))}{n}+\frac{(2p-1)b_1}{n^2}-\frac{p^2e_1}{n^2}+\mathcal{O}\big(\frac{1}{n^3}\big) \tag{A.21}$$

where e_1 is defined in Lemma 3.2-(d).

A.5. **Study of $I_2^{(2,p)}$.** We get from (4):

$$2\big(F_n(t)-G_n^{(2,p)}(t)\big)F(t)=2(F_n(t)-G_n^{(2,p)}(t)F(t)\,\mathbb{I}_{[X_1^*,X_n^*)}(t) + \frac{\big(-2t-2(1-p)X_2^*+2(2-p)X_1^*\big)F(t)}{n(X_2^*-X_1^*)}\,\mathbb{I}_{[(2-p)X_1^*-(1-p)X_2^*,X_1^*)}(t) + \frac{\big(-2t+2(1+p)X_n^*-2pX_{n-1}^*\big)F(t)}{n(X_n^*-X_{n-1}^*)}\,\mathbb{I}_{[X_n^*,(1+p)X_n^*-pX_{n-1}^*)}(t).$$

Using again the notation K_0 and K_1 and Taylor formula with integral remainder, the new terms to control are

$$-\mathbb{E}\Big(\frac{(1-p)^2(X_2^*-X_1^*)}{n}F\big((2-p)X_1^*-(1-p)X_2^*\big)\Big)+\mathbb{E}\Big(\frac{p^2(X_n^*-X_{n-1}^*)}{n}F(X_n^*)\Big)$$
$$+\mathbb{E}\Big(\frac{1}{n(X_2^*-X_1^*)}\int_{(2-p)X_1^*-(1-p)X_2^*}^{X_1^*}f(t)(X_1^*-t)\big(X_1^*-t-2(1-p)(X_2^*-X_1^*)\big)\,\mathrm{d}t\Big)$$
$$+\mathbb{E}\Big(\frac{1}{n(X_n^*-X_{n-1}^*)}\int_{X_n^*}^{(p+1)X_n^*-pX_{n-1}^*}f(t)\big(t-(1+p)X_n^*+pX_{n-1}^*\big)^2\,\mathrm{d}t\Big).$$

With Lemma 3.2, we may conclude that

$$2\,\mathbb{E}\int_{(2-p)X_1^*-(1-p)X_2^*}^{(p+1)X_n^*-pX_{n-1}^*}\big(F_n(t)-G_n^{(2,p)}\big)F(t)\,\mathrm{d}t=\frac{p^2e_1}{n^2}-\frac{(2p-1)b_1}{n^2} + \frac{(2p-1)(1-\mathbb{E}\,(X_1))}{n}-\frac{1}{3n^2}+\mathcal{O}\big(\frac{1}{n^3}\big). \tag{A.22}$$

Details are omitted.

A.6. **Conclusion for $G_n^{(2,p)}$.** Collecting the results obtained in (A.9)-(A.10) and (A.21)-(A.22), we obtain that

$$2\,\mathbb{E}\int_0^1\big(G_n^{(2,p)}(t)-F_n(t)\big)\big(F_n(t)-F(t)\big)\,\mathrm{d}t=-\frac{1}{3n^2}+\mathcal{O}(\frac{1}{n^3}),$$

so Proposition 3.5 is proved also for $G_n^{(2,p)}$.

Appendix Appendix B. Parameters of the 16 normal mixtures densities

Number	Name	Distribution function: $\sum_{\ell=0}^{k} w_\ell \mathcal{N}(\mu_\ell, \sigma_\ell^2)$
1	Normal	$\mathcal{N}(0,1)$
2	Skewed unimodal	$\frac{1}{5}\mathcal{N}(0,1) + \frac{1}{5}\mathcal{N}(\frac{1}{2}, (\frac{2}{3})^2) + \frac{3}{5}\mathcal{N}(\frac{13}{12}, (\frac{5}{9})^2)$
3	Strongly skewed	$\sum_{\ell=0}^{7} \frac{1}{8}\mathcal{N}(3((\frac{2}{3})^\ell - 1), (\frac{2}{3})^{2\ell})$
4	Kurtotic unimodal	$\frac{2}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(0, (\frac{1}{10})^2)$
5	Outlier	$\frac{1}{10}\mathcal{N}(0,1) + \frac{9}{10}\mathcal{N}(0, (\frac{1}{10})^2)$
6	Bimodal	$\frac{1}{2}\mathcal{N}(-1, (\frac{2}{3})^2) + \frac{1}{2}\mathcal{N}(1, (\frac{2}{3})^2)$
7	Separated bimodal	$\frac{1}{2}\mathcal{N}(-\frac{3}{2}, (\frac{1}{2})^2) + \frac{1}{2}\mathcal{N}(\frac{3}{2}, (\frac{1}{2})^2)$
8	Asymmetric bimodal	$\frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}(\frac{3}{2}, (\frac{1}{3})^2)$
9	Trimodal	$\frac{9}{20}\mathcal{N}(-\frac{6}{5}, (\frac{3}{5})^2) + \frac{9}{20}\mathcal{N}(\frac{6}{5}, (\frac{3}{5})^2) + \frac{1}{10}\mathcal{N}(0, (\frac{1}{4})^2)$
10	Claw	$\frac{1}{2}\mathcal{N}(0,1) + \sum_{\ell=0}^{4} \frac{1}{10}\mathcal{N}(\frac{\ell}{2} - 1, (\frac{1}{10})^2)$
11	Double claw	$\frac{49}{100}\mathcal{N}(-1, (\frac{2}{3})^2) + \frac{49}{100}\mathcal{N}(1, (\frac{2}{3})^2) + \sum_{\ell=0}^{6} \frac{1}{350}\mathcal{N}(\frac{\ell-3}{2}, (\frac{1}{100})^2)$
12	Asymmetric claw	$\frac{1}{2}\mathcal{N}(0,1) + \sum_{\ell=-2}^{2} \frac{2^{1-\ell}}{31}\mathcal{N}(\ell + \frac{1}{2}, (\frac{2^{-\ell}}{10})^2)$
13	Asymmetric double claw	$\sum_{\ell=0}^{1} \frac{46}{100}\mathcal{N}(2\ell - 1, (\frac{2}{3})^2) + \sum_{\ell=1}^{3} \frac{1}{300}\mathcal{N}(-\frac{\ell}{2}, (\frac{1}{100})^2) + \sum_{\ell=1}^{3} \frac{7}{300}\mathcal{N}(\frac{\ell}{2}, (\frac{7}{100})^2)$
14	Smooth comb	$\sum_{\ell=0}^{5} \frac{2^{5-\ell}}{63}\mathcal{N}(\frac{65-96(1/2)^\ell}{21}, \frac{(32/63)^2}{2^{2\ell}})$
15	Discrete comb	$\sum_{\ell=0}^{2} \frac{2}{7}\mathcal{N}(\frac{12\ell-15}{7}, (\frac{2}{7})^2) + \sum_{\ell=8}^{10} \frac{1}{21}\mathcal{N}(\frac{2\ell}{7}, (\frac{1}{21})^2)$
16	Distant bimodal	$\frac{1}{2}\mathcal{N}(-\frac{5}{2}, (\frac{1}{6})^2) + \frac{1}{2}\mathcal{N}(\frac{5}{2}, (\frac{1}{6})^2)$

Table 1. Distribution functions used in the simulation study: #1-#15 are from Marron and Wand (1992), #16 from Janssen et al. (1995)

Avignon University, LMA EA2151, Campus Jean-Henri Fabre, 301 rue Baruch de Spinoza, BP 21239, F-84916 Avignon Cedex 9, France
Email address: delphine.blanke@univ-avignon.fr

Sorbonne Universités, UPMC Univ Paris 06, Laboratoire de Probabilités, Statistique et Modélisation, LPSM, 4 place Jussieu, F-75005, Paris, France
Email address: denis.bosq@upmc.fr