Piecewise linear continuous estimators of the quantile function

Delphine Blanke, Denis Bosq

PIECEWISE LINEAR CONTINUOUS ESTIMATORS OF THE QUANTILE FUNCTION

DELPHINE BLANKE AND DENIS BOSQ

Abstract. In Blanke and Bosq (2018), families of piecewise linear estimators of the distribution function F were introduced. It was shown that they reduce the mean integrated squared error (MISE) of the empirical distribution function F_n and that the minimal MISE was reached by connecting the midpoints $(\frac{X_k^*+X_{k+1}^*}{2}, \frac{k}{n})$, with $X_1^*, \ldots, X_n^*$ the order statistics. In this contribution, we consider the reciprocal estimators, built respectively for known and unknown support of distribution, for estimating the quantile function F^{-1}. We prove that these piecewise linear continuous estimators again strictly improve the MISE of the classical sample quantile function F_n^{-1}.

1. Introduction

If $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) real random variables with absolutely continuous distribution function F, the quantile function is defined as $F^{-1}(t) = \inf\{x : F(x) \geq t\}$. The sample (or empirical) quantile function is then $F_n^{-1}(t) = \inf\{x : F_n(x) \geq t\}$, with $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{]-\infty,x]}(X_i)$, $x \in \mathbb{R}$ and $\mathbb{I}_A$ denotes the indicator function of the set A. This is equivalent to $F_n^{-1}(t) = X_k^*$ for $t \in]\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \ldots, n$ and where $X_1^* < \cdots < X_n^*$ (almost surely) denotes the ordered sample. We study the properties of two piecewise linear alternatives of F_n^{-1} that respectively address the cases of known and unknown support of the density f. Actually, these estimators are the reciprocals of two particular estimators considered in Blanke and Bosq (2018) to estimate F. More precisely, in this last cited reference, the MISE of a general family of polygonal estimators of the distribution function is studied. These estimators consist in linearly interpolating F_n at different points, namely $(X_k^* + p(X_{k+1}^* - X_k^*), \frac{k}{n})$, $k = 1, \ldots, n-1$, where p is a chosen parameter in [0,1]. For example, $p = 0$ corresponds to a piecewise linear interpolation at $(X_k^*, \frac{k}{n})$, $p = 1$ at $(X_{k+1}^*, \frac{k}{n})$, and the choice $p = \frac{1}{2}$ connects the midpoints $(\frac{X_k^*+X_{k+1}^*}{2}, \frac{k}{n})$. It is then shown in Blanke and Bosq (2018) that for all p chosen in]0,1[, the MISE of F_n is strictly improved and that it is minimal at $p = \frac{1}{2}$ (while the choices $p = 0$ or 1 cannot be recommended). The reciprocals of estimators connecting the midpoints of F_n are studied in this contribution to estimate the quantile function, they join the midpoints of F_n^{-1}, and their formulation depends on whether or not the support is known.

A large literature exists on quantile estimation or L-statistics (linear functions of order statistics): we may refer to the review proposed by Poiraud-Casanova and Thomas-Agnan (1998) or to the detailed introductions in Sheather and Marron (1990) and Zelterman (1990) for smooth quantile estimation, and to Cheng and Parzen (1997) for an unified approach. The origin of the quantile estimators studied in this paper can go back to Hazen (1914) in hydrology (see Harter, 1984, for a discussion about plotting positions). Even if their good behaviour had been outlined by Parzen (1979); Dielman et al. (1994), as far as we can judge, there has been no theoretical study of their statistical behavior until now.

The paper is organized as follows. In Section 2, we introduce our two piecewise quantile estimators and give their first properties deduced from their proximity to the sample quantile function F_n^{-1}. The main result of this paper is the derivation of their MISE established in Theorem 2.4. It appears that the piecewise quantile estimators strictly improve the sample quantile function and have an equivalent MISE up to the second order. A conclusion and discussion about possible extensions of our results appear in Section 3. Finally, the proof of the most technical results are postponed to the appendix.

2. The piecewise quantile estimators

2.1. **Definition.** For independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with compact support $[a, b]$ and absolutely continuous distribution function F, we introduce two

continuous piecewise linear estimators of the quantile function F^{-1}. These estimators are the reciprocals of the two estimators of F, studied in Blanke and Bosq (2018), which linearly interpolate the empirical cumulative distribution function F_n at its midpoints, and which minimise the MISE among the set of polygonal estimators considered in the latter reference.

Our first quantile estimator, G_{n1}^{-1}, addresses the case of a known support $[a, b]$ by using this support in its construction. The second estimator, G_{n2}^{-1}, modifies G_{n1}^{-1} at its both ends, and handles by this way the case of an unknown support. Note that even if the results of this article are established for distributions with compact support, the definition of G_{n2}^{-1} is adapted for the case where no information on the support of the distribution is available (and so can even be infinite).

Definition 2.1. *1) For known support $[a, b]$, we define*

$$G_{n1}^{-1}(t) = \begin{cases} 2nt(X_1^* - a) + a & \text{if } t \in [0, \frac{1}{2n}], \\ (nt - k + \frac{1}{2})(X_{k+1}^* - X_k^*) + X_k^* & \text{if } t \in]\frac{2k-1}{2n}, \frac{2k+1}{2n}],\ k = 1, \ldots, n-1, \\ b - 2n(1-t)(b - X_n^*) & \text{if } t \in]1 - \frac{1}{2n}, 1]. \end{cases} \tag{1}$$

2) In the general case, we set $G_{n2}^{-1}(t) = G_{n1}^{-1}(t)$ for $t \in]\frac{1}{2n}, 1 - \frac{1}{2n}]$, and for $n \geq 2$,

$$G_{n2}^{-1}(t) = \begin{cases} (nt - \frac{1}{2})(X_2^* - X_1^*) + X_1^* & \text{if } t \in [0, \frac{1}{2n}], \\ (nt - n + \frac{1}{2})(X_n^* - X_{n-1}^*) + X_n^* & \text{if } t \in]1 - \frac{1}{2n}, 1]. \end{cases} \tag{2}$$

Let us recall that the classical sample quantile function is the generalized inverse function of F_n defined by

$$F_n^{-1}(t) = \inf\{x \ :\ F_n(x) \geq t\}$$

and is equivalent to $F_n^{-1}(t) = X_k^*$ for $t \in]\frac{k-1}{n}, \frac{k}{n}]$, $k = 1, \ldots, n$. Our estimators simply regularize F_n^{-1} by connecting its midpoints on $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ and are extended in a natural way at both ends (towards the support for G_{n1}^{-1} and by lengthening the last segments for G_{n2}^{-1}).

Let us notice that connecting the midpoints of F_n^{-1} on $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ is an old proposition, suggested in Hazen (1914), which remains popular and used in hydrology. Such an estimator also appears in Harter (1984); Parzen (1979); Parrish (1990); Dielman et al. (1994) and is implemented in statistical packages (Hyndman and Fan, 1996). But according to these authors, even with good performance in simulations and good properties of construction, it presents several problems:

- not being justified on the basis of an estimation argument (Hyndman and Fan, 1996),
- being restricted on the support $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ (Dielman et al., 1994),
- and only suited to symmetric distributions (Parzen, 1979; Dielman et al., 1994).

The results presented in this contribution address the above-mentioned drawbacks. We establish the asymptotic behavior of the estimators G_{n1}^{-1} and G_{n2}^{-1} defined on [0,1], and we show that they are always better than the sample quantile function in terms of MISE.

2.2. **First properties.** First, note that estimators G_{nj}^{-1}, $j = 1, 2$, are exemples of linear functions of order statistics. Such L-estimators have been extensively studied and share natural properties expected for the quantile function. They are defined as weighted averages of consecutive order statistics. Those of them involving one or two order statistics can be written as: $(1-\gamma)X_k^* + \gamma X_{k+1}^*$, where $(k - \ell)/n \leq t < (k - \ell + 1)/n$ and $\gamma = nt + \ell - k$ with $\ell \in \mathbb{R}$ a constant determined by the considered estimator (see Hyndman and Fan, 1996, for values taken by ℓ according to the chosen sample quantile). In our case, the choice $\ell = \frac{1}{2}$ gives $G_{nj}^{-1}(t)$, $j = 1, 2$, for $t \in [\frac{1}{2n}, 1 - \frac{1}{2n}]$. For the intervals $[0, \frac{1}{2n}]$ and $[1 - \frac{1}{2n}, 1]$, the same expression holds for G_{n2}^{-1} by setting $k = 1$ and $k = n - 1$ respectively on the valid definition over $[\frac{1}{2n}, 1 - \frac{1}{2n}]$. The following proposition reviews some of other natural properties of our estimators.

Proposition 2.1. *For $j = 1, 2$, we get that the estimators G_{nj}^{-1} are*

(a) continuous on [0,1],
(b) symmetric,
(c) invariant by translation (only on $[\frac{1}{2n}, 1 - \frac{1}{2n}]$ for G_{n1}^{-1}),
(d) equal to the usual sample median for $t = \frac{1}{2}$.

Proof. (a) Clear by construction.

(b) We have to check that $G^{-1}_{-X,nj}(t) = -G^{-1}_{nj}(1-t)$ for $G^{-1}_{-X,nj}$ built with $(-X_1,\dots,-X_n)$. Symmetry is obtained by substituting X^*_k by $-X^*_{n-k+1}$ for $k=1,\dots,n$ and $[a,b]$ by $[-b,-a]$ in (1)-(2).

(c) For $Y_k = X_k + c$, $k=1,\dots,n$ with some constant c, we have to establish that $G^{-1}_{Y,nj}(t) = G^{-1}_{nj}(t) + c$ if $G^{-1}_{Y,nj}$ is the sample quantile estimator built with $Y_1,\dots,Y_n$. The result is clear with $Y^*_k = X^*_k + c$ for all $k=1,\dots,n$ in (1)-(2). The property is no longer true for $G^{-1}_{n1}(t)$ with $t \in [0, \frac{1}{2n}]$ or $t \in [1-\frac{1}{2n}, 1]$.

(d) For $j=1,2$ and $n=1$, $G^{-1}_{nj}(\frac{1}{2}) = X^*_1$. For $n \geq 2$ and $n=2p$, $G^{-1}_{nj}(\frac{1}{2}) = \frac{X^*_p + X^*_{p+1}}{2}$ while for $n=2p+1$, $G^{-1}_{nj}(\frac{1}{2}) = X^*_{p+1}$.

□ □

The next immediate lemma specifies the proximity between F^{-1}_n and G^{-1}_{nj} and will be useful for establishing the convergence of our estimators. Note that frow now on, we set $a=0$ and $b=1$ to simplify the presentation of the results.

Lemma 2.2. *1) For $j=1,2$ and $k=1,\dots,n-1$,*

$$G^{-1}_{nj}(t) - F^{-1}_n(t) = \begin{cases} (nt-k+\frac{1}{2})(X^*_{k+1}-X^*_k) & \text{if } t \in]\frac{k}{n}-\frac{1}{2n}, \frac{k}{n}] \\ (nt-k-\frac{1}{2})(X^*_{k+1}-X^*_k) & \text{if } t \in]\frac{k}{n}, \frac{k}{n}+\frac{1}{2n}]. \end{cases}$$

2) For $t \in [0, \frac{1}{2n}]$, $G^{-1}_{n1}(t) - F^{-1}_n(t) = (2nt-1)X^*_1$ while $G^{-1}_{n2}(t) - F^{-1}_n(t) = (nt-\frac{1}{2})(X^*_2 - X^*_1)$.

3) For $t \in [1-\frac{1}{2n}, 1]$, $G^{-1}_{n1}(t) - F^{-1}_n(t) = (2nt-2n+1)(1-X^*_n)$ while $G^{-1}_{n2}(t) - F^{-1}_n(t) = (nt-n+\frac{1}{2})(X^*_n - X^*_{n-1})$.

Corollary 2.3. *If F is absolutely continuous with density f such that f is continuous on [0,1] and $\inf_{x\in[0,1]} f(x) \geq c_0$ for some positive constant c_0, one obtains that $\sup_{t\in[0,1]} \big|G^{-1}_{nj}(t) - F^{-1}_n(t)\big| = \mathcal{O}_p(n^{-1})$, $j=1,2$.*

Proof. First recall that the joint density of (X^*_k, X^*_{k+1}) (see e.g. David and Nagaraja, 2003, p. 12) is given by

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

Next, integrations by parts imply that $\mathbb{E}\,(X^*_{k+1} - X^*_k) = C^k_n \int_0^1 F^k(x)(1-F(x))^{n-k}\,\mathrm{d}x$ so that

$$\mathbb{E}\,(X^*_{k+1} - X^*_k) = C^k_n \int_0^1 u^k (1-u)^{n-k} \frac{1}{f(F^{-1}(u))}\,\mathrm{d}u \leq \frac{C^k_n}{c_0} \int_0^1 u^k (1-u)^{n-k}\,\mathrm{d}u.$$

From the standard result $\int_0^1 u^k(1-u)^{n-k}\,\mathrm{d}u = \frac{k!(n-k)!}{(n+1)!}$, we may deduce that $\mathbb{E}\,(X^*_{k+1} - X^*_k) = \mathcal{O}(\frac{1}{n+1})$ uniformly in k. One easily concludes with Lemma 2.2 and Lemma 3.2(a)-(b) in Blanke and Bosq (2018) (recalled in the Appendix, see Lemma 3.1) together with Markov inequality. □ □

We may deduce that the two estimators are asymptotically equivalent to F^{-1}_n; for example, they get the same limit in distribution since $\sqrt{n}(G^{-1}_{nj}(t) - F^{-1}_n(t)) \xrightarrow[n\to\infty]{p} 0$ for $j=1,2$.

2.3. **Mean integrated squared error.** We now give the main result of this contribution showing that the estimators strictly improve the sample quantile function in terms of MISE and are equivalent up to its second order.

Theorem 2.4. *If F is absolutely continuous with density f such that f is C^1 on [0,1] and $\inf_{x\in[0,1]} f(x) > 0$, we get that for, $j=1,2$,*

$$\int_0^1 \mathbb{E}\left(G^{-1}_{nj}(t) - F^{-1}(t)\right)^2 \mathrm{d}t \;=\; \int_0^1 \mathbb{E}\left(F^{-1}_n(t) - F^{-1}(t)\right)^2 \mathrm{d}t - \frac{1}{4n^2} \int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-\frac{5}{2}}\big).$$

The proof of Theorem 2.4 is based on the decomposition of $\big(G^{-1}_{nj}(t) - F^{-1}(t)\big)^2$ into

$$\big(G^{-1}_{nj}(t) - F^{-1}_n(t)\big)^2 + \big(F^{-1}_n(t) - F^{-1}(t)\big)^2 + 2\big(G^{-1}_{nj}(t) - F^{-1}_n(t)\big)\big(F^{-1}_n(t) - F^{-1}(t)\big)$$

with the following proposition proved in the appendix.

Proposition 2.5. *Under the assumptions of Theorem 2.4, we obtain*

1) for $j = 1, 2$,

$$\mathbb{E}\int_0^1 \big(G_{nj}^{-1}(t) - F_n^{-1}(t)\big)^2 \,\mathrm{d}t = \frac{1}{6n(n+1)}\int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-3}\big).$$

2) for $j = 1$,

$$\mathbb{E}\int_0^1 \big(G_{n1}^{-1}(t) - F_n^{-1}(t)\big)F_n^{-1}(t)\,\mathrm{d}t = \frac{\mathbb{E}\,(1-X_n^*)}{4n} - \frac{1}{4n(n+1)}\int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-3}\big)$$

while, for $j = 2$,

$$\mathbb{E}\int_0^1 \big(G_{n2}^{-1}(t) - F_n^{-1}(t)\big)F_n^{-1}(t)\,\mathrm{d}t = \frac{\mathbb{E}\,(X_n^* - X_{n-1}^*)}{8n} - \frac{1}{4n(n+1)}\int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-3}\big).$$

3) for $j = 1$,

$$\mathbb{E}\int_0^1 \big(G_{n1}^{-1}(t) - F_n^{-1}(t)\big)F^{-1}(t)\,\mathrm{d}t = \frac{\mathbb{E}\,(1-X_n^*)}{4n} - \frac{1}{24n^2}\int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-\frac{5}{2}}\big)$$

while, for $j = 2$,

$$\mathbb{E}\int_0^1 \big(G_{n2}^{-1}(t) - F_n^{-1}(t)\big)F^{-1}(t)\,\mathrm{d}t = \frac{\mathbb{E}\,(X_n^* - X_{n-1}^*)}{8n} - \frac{1}{24n^2}\int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-\frac{5}{2}}\big).$$

Moreover, in the proof of Proposition 2.5, the following result is established, see equation (7) in the appendix. We highlight this result because it might be useful for other applications.

Lemma 2.6. *If F is absolutely continuous with density f such that f is C^1 on [0,1] and $\inf_{x\in[0,1]} f(x) > 0$, we get that,*

$$\int_0^1\int_x^1 \big(1 - F(y) + F(x)\big)^n \,\mathrm{d}y\,\mathrm{d}x = \frac{1}{n+1}\int_0^1 \frac{1}{f(y)}\,\mathrm{d}y - \frac{1}{2(n+1)(n+2)f^2(1)} - \frac{1}{2(n+1)(n+2)f^2(0)} + \mathcal{O}(n^{-3}).$$

Note that Theorem 2.4 illustrates the classical phenomenon of deficiency with the dominant term given by the MISE of the sample quantile function. This phenomenon appears also for kernel quantile estimators, see Falk (1984); Sheather and Marron (1990), as well as for estimators inverting kernel estimators of the distribution function, see Azzalini (1981). In these works, an optimal choice of the bandwidth allows a gain compared to the MISE of F_n that is a $\mathcal{O}(n^{-\frac{4}{3}})$ for the term of the second order. By this way, these estimators are more efficient than our piecewise linear ones, that have a gain of only $\mathcal{O}(n^{-2})$, but G_{nj}^{-1}, $j = 1, 2$, present the immediate advantages to not depend on any smoothing parameter and can be plotted directly in an easy way. As indicated in Proposition 2.1, they also meet the qualities expected for empirical quantiles (see also Hyndman and Fan, 1996).

To conclude this part, we complete Theorem 2.4 with the MISE of the sample quantile function (in accordance with the Bahadur representation) as we did not find the explicit result in the literature.

Proposition 2.7. *Under the assumptions of Theorem 2.4, we have*

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

If we suppose moreover that f is C^2 on [0,1], we get that

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

Proof. We start from

$$\int_0^1 \big(F_n^{-1}(t) - F^{-1}(t)\big)^2 \,\mathrm{d}t = \sum_{k=1}^n \int_{\frac{k-1}{n}}^{\frac{k}{n}} \big(X_k^* - F^{-1}(t)\big)^2 \,\mathrm{d}t$$
$$= \frac{1}{n}\sum_{k=1}^n X_k^{*2} + \mathbb{E}\,(X_1^2) - 2\sum_{k=1}^n X_k^* \int_{\frac{k-1}{n}}^{\frac{k}{n}} F^{-1}(t)\,\mathrm{d}t$$

so that $\mathbb{E}\int_0^1 \left(F_n^{-1}(t)-F^{-1}(t)\right)^2 \mathrm{d}t = 2\big(\mathbb{E}\,(X_1^2) - \sum_{k=1}^n \mathbb{E}\,(X_k^*)\int_{\frac{k-1}{n}}^{\frac{k}{n}} F^{-1}(t)\,\mathrm{d}t\big)$. The main task is the evaluation of the last term. Taylor formula and continuity of f' give that uniformly over k,

$$\int_{\frac{k-1}{n}}^{\frac{k}{n}} F^{-1}(t)\,\mathrm{d}t = \frac{1}{n}F^{-1}(\frac{k-1}{n}) + \frac{1}{2n^2}\frac{1}{f(F^{-1}(\frac{k-1}{n}))} + \mathcal{O}(n^{-3})$$

which in turn gives a $\mathcal{O}(n^{-2})$ for the term of rest after multiplying it by $\sum_{k=1}^n \mathbb{E}\,(X_k^*) = n\mathbb{E}\,(X_1)$. Next, again using Taylor formula to control the remaining terms, we may write

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^n \mathbb{E}\,(X_k^*)F^{-1}(\frac{k-1}{n}) \\
&= \frac{1}{n}\sum_{k=1}^n \mathbb{E}\,(X_k^*)\Big[F^{-1}(\frac{k}{n+1}) - \frac{n-(k-1)}{n(n+1)}\frac{1}{f(F^{-1}(\frac{k-1}{n}))}\Big] + \mathcal{O}(n^{-2}) \\
&= \frac{1}{n}\sum_{k=1}^n \mathbb{E}\,(X_k^*)\Big[F^{-1}(\frac{k}{n+1}) - \frac{1}{nf(F^{-1}(\frac{k}{n+1}))} + \frac{k}{n(n+1)f(F^{-1}(\frac{k}{n+1}))}\Big] + \mathcal{O}(n^{-2}). \qquad (3)
\end{aligned}$$

Next, we apply results concerning the expectation of linear combinations of order statistics, $\frac{1}{n}\sum_{k=1}^n J(\frac{k}{n+1})X_k^*$, given in Stigler (1974) and Helmers (1980). First, since F^{-1} is twice differentiable on [0,1], we may adapt equation (5.11) in the proof of Theorem 2.2 in Helmers (1980) to obtain that

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^n \mathbb{E}\,(X_k^*)F^{-1}(\frac{k}{n+1}) &= \int_0^1 (F^{-1}(t))^2\,\mathrm{d}t \\
&\quad - \frac{1}{2n}\int_0^1 \frac{t(1-t)}{f^2(F^{-1}(t))}\,\mathrm{d}t + \frac{1}{n}\int_0^1 (\frac{1}{2}-t)\frac{F^{-1}(t)}{f(F^{-1}(t))}\,\mathrm{d}t + \mathcal{O}(n^{-\frac{3}{2}}). \qquad (4)
\end{aligned}$$

For the two last terms in (3) involving the density f, we may apply the Theorem 4 of Stigler (1974) that does not require the existence of f''. This allows to obtain that

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}\,(X_k^*)\frac{1}{f(F^{-1}(\frac{k}{n+1}))} = \int_0^1 \frac{F^{-1}(t)}{f(F^{-1}(t))}\,\mathrm{d}t + o(n^{-\frac{1}{2}}) \qquad (5)$$

and

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}\,(X_k^*)\frac{k}{n+1}\frac{1}{f(F^{-1}(\frac{k}{n+1}))} = \int_0^1 \frac{tF^{-1}(t)}{f(F^{-1}(t))}\,\mathrm{d}t + o(n^{-\frac{1}{2}}). \qquad (6)$$

To conclude the proof, one may note that if f'' exists and is continuous, one may apply the Helmers (1980)'s result to get a $\mathcal{O}(n^{-2})$ instead of $\mathcal{O}(n^{-\frac{3}{2}})$ in (4) and a $\mathcal{O}(n^{-1})$ instead of a $o(n^{-\frac{1}{2}})$ in (5)-(6). □ □

3. Discussion

We have studied two smoothed quantile estimators, G_{n1}^{-1} and G_{n2}^{-1} 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 are simple generalizations of the quantile process $F_n^{-1}(t)$, they do not depend on any smoothing parameter and can be plotted directly without any computation. They also meet the standard properties expected for quantiles. Our main result points ahead that they strictly improve the MISE of F_n^{-1}. Moreover, the two estimators have equivalent MISE up to the order n^{-2}. The first one uses the support of the density in its construction while the second one does not require the knowledge of this support. Even if their good numerical properties had already been pointed out in the literature, see for example Parrish (1990); Dielman et al. (1994), their theoretical study had not been carried out until now. We hope that this paper can give additional motivation for their study and their use in practical problems.

Our results are in agreement with those established for equivalent estimators of the distribution function in Blanke and Bosq (2018). In this reference, general families of estimators of F have been considered, by joining linearly the empirical distribution function F_n at some defined points. It is established that estimators joining the midpoints (as their reciprocal forms studied in the present article for quantile estimation) reach a minimal MISE at the order n^{-2}. In addition, it is shown that piecewise linear estimators joining the order statistics (either the points $(X_k^*, \frac{k}{n})$ or $(X_{k+1}^*, \frac{k}{n})$) do not

improved the MISE of F_n at this order. It would be interesting to see if this bad behavior can also be established also for quantile estimation since the simple kernel estimator (joining two consecutive order statistics) is still popular among practitioners. Indeed, more general sample quantiles of the type $Q_n(t) = (1-\gamma)X_k^* + \gamma X_{k+1}^*$ when $(k-\ell)/n \leq t < (k-\ell+1)/n$ for some ℓ, could be studied with the techniques of our article in order to compare their asymptotic behaviors (work in progress).

Various other extensions of our 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. As noted previously, the estimator G_{n2}^{-1} seems naturally suitable for such a framework. We may also remark, see Babu et al. (2002), that 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. It should be interesting to look at the transformation of our estimators in these cases. Also, some numerical studies on F_n^{-1}, G_{n1}^{-1} and G_{n2}^{-1}, not exposed here, have been conducted for gaussian mixtures (in the same way as in Blanke and Bosq, 2018) and give good results even in these unbounded cases.

Finally, conditional quantile estimation is now a large field of research; we can refer to the nice survey of Poiraud-Casanova and Thomas-Agnan (1998) or to the more recent handbooks of Koenker (2005); Koenker et al. (2018). It allows to take into account the influence of covariates on the studied distribution and has multiple applications in medicine, economics and finance. Also, they represent a robust alternative to the conditional mean and are involved in curve estimation, see e.g. Aragon et al. (2005) for frontier estimation and Leconte et al. (2002) for random censorship. It would be interesting to see how our estimators could be written in these frameworks, and see if their easy implementation could offer an interesting alternative to existing usual estimators.

References

Y. Aragon, A. Daouia, and C. Thomas-Agnan. Nonparametric frontier estimation: a conditional quantile-based approach. *Econometric Theory*, 21(2):358–389, 2005.

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.

D. Blanke and D. Bosq. Polygonal smoothing of the empirical distribution function. *Stat. Inference Stoch. Process.*, 21(2):263–287, 2018.

C. Cheng and E. Parzen. Unified estimators of smooth quantile and quantile density functions. *J. Statist. Plann. Inference*, 59(2):291–307, 1997.

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.

T. Dielman, C. Lowry, and R. Pfaffenberger. A comparison of quantile estimators. *Communications in Statistics - Simulation and Computation*, 23(2):355–371, 1994.

M. Falk. Relative deficiency of kernel type estimators of quantiles. *Ann. Statist.*, 12(1):261–268, 1984.

H. L. Harter. Another look at plotting positions. *Communications in Statistics - Theory and Methods*, 13(13):1613–1633, 1984.

A. Hazen. Storage to be providing in impounding reservoirs for municipal water supply (with discussion). *Transaction of the American society of civil engineers*, 77:1539–1669, 1914.

R. Helmers. Edgeworth expansions for linear combinations of order statistics with smooth weight functions. *Ann. Statist.*, 8(6):1361–1374, 1980.

R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50 (4):361–365, 1996.

R. Koenker. *Quantile Regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, New York, 2005.

R. Koenker, V. Chernozhukov, X. He, and L. Peng, editors. *Handbook of quantile regression*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2018.

E. Leconte, S. Poiraud-Casanova, and C. Thomas-Agnan. Smooth conditional distribution function and quantiles under random censorship. *Lifetime Data Anal.*, 8(3):229–246, 2002.

R. S. Parrish. Comparison of quantile estimators in normal sampling. *Biometrics*, 46(1):247–257, 1990.
E. Parzen. Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.*, 74(365):105–131, 1979.
S. Poiraud-Casanova and C. Thomas-Agnan. Quantiles conditionnels. *Journal de la société française de statistique*, 139(4):31–44, 1998.
S. J. Sheather and J. S. Marron. Kernel quantile estimators. *J. Amer. Statist. Assoc.*, 85(410): 410–416, 1990.
S. M. Stigler. Linear functions of order statistics with smooth weight functions. *Ann. Statist.*, 2: 676–693, 1974.
D. Zelterman. Smooth nonparametric estimation of the quantile function. *J. Statist. Plann. Inference*, 26(3):339–352, 1990.

Appendix

The two following lemmas are useful for calculations.

Lemma 3.1 (Blanke and Bosq, 2018, Lemma 3.2)**.** *If f is continuous on [0,1] and $\inf_{x\in[0,1]} f(x) \geq c_0$ for some positive constant c_0 then, for all integers $r \geq 0$ and $m \geq 1$, not depending on n, we get*

(a)

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

(b)

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

(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)

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

Lemma 3.2 (Blanke and Bosq, 2018, Proposition A1)**.** *If h is measurable and integrable on $[0,1]^2$, then*

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

Proof of Proposition 2.5. 1) We start from Lemma 2.2 and simple integrations give

$$\mathbb{E}\int_0^1 \left(G_{n1}^{-1}(t) - F_n^{-1}(t)\right)^2 \mathrm{d}t = \frac{\mathbb{E}\left(X_1^*\right)^2}{6n} + \frac{\mathbb{E}\left(1-X_n^*\right)^2}{6n} + \frac{\sum_{k=1}^{n-1}\mathbb{E}\left(X_{k+1}^* - X_k^*\right)^2}{12n}$$

for $j = 1$, while for $j = 2$, one gets

$$\mathbb{E}\int_0^1 \left(G_{n2}^{-1}(t) - F_n^{-1}(t)\right)^2 \mathrm{d}t = \frac{\mathbb{E}\left(X_2^* - X_1^*\right)^2}{24n} + \frac{\mathbb{E}\left(X_n^* - X_{n-1}^*\right)^2}{24n} + \frac{\sum_{k=1}^{n-1}\mathbb{E}\left(X_{k+1}^* - X_k^*\right)^2}{12n}.$$

Lemma 3.1 implies that the two first terms in these expressions are negligible in $\mathcal{O}\big(n^{-3}\big)$. Lemma 3.2 implies that

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

Next, integrations by parts give that

$$\sum_{k=1}^{n-1} \mathbb{E}\left(X_{k+1}^* - X_k^*\right)^2 = -2\int_0^1 y\mathbb{P}(X_1^* > y)\,\mathrm{d}y - 2\int_0^1 (1-x)\mathbb{P}(X_n^* \leq x)\,\mathrm{d}x + 2\int_0^1\int_x^1 \big(1-F(y)+F(x)\big)^n\,\mathrm{d}y\,\mathrm{d}x.$$

Setting $t = y^2$ and $t = (1-x)^2$ in the two first integrals give a $\mathcal{O}(n^{-2})$ for these terms with Lemma 3.1. For the term $2\int_0^1\int_x^1 \big(1-F(y)+F(x)\big)^n \,\mathrm{d}y\,\mathrm{d}x$, we perform the change of variables $y = F^{-1}(t)$, $x = F^{-1}(s)$ to get

$$\int_0^1\int_x^1 \big(1-F(y)+F(x)\big)^n \,\mathrm{d}y\,\mathrm{d}x = \int_0^1\int_s^1 (1-t+s)^n \frac{1}{f(F^{-1}(t))}\frac{1}{f(F^{-1}(s))}\,\mathrm{d}s\,\mathrm{d}t.$$

Again multiple integrations by parts lead to

$$\int_0^1\int_x^1 \big(1-F(y)+F(x)\big)^n \,\mathrm{d}y\,\mathrm{d}x = \frac{1}{n+1}\int_0^1 \frac{1}{f(y)}\,\mathrm{d}y - \frac{1}{2(n+1)(n+2)f^2(1)} - \frac{1}{2(n+1)(n+2)f^2(0)} + \mathcal{O}(n^{-3}). \quad (7)$$

Now, one may conclude that

$$\sum_{k=1}^{n-1}\mathbb{E}\left(X^*_{k+1}-X^*_k\right)^2 = \frac{2}{n+1}\int_0^1 \frac{1}{f(x)}\,\mathrm{d}x + \mathcal{O}\big(n^{-2}\big) \quad (8)$$

and the result follows.

2) From Lemma 2.2 and $F_n^{-1}(t) = X^*_k$ for $t \in]\frac{k-1}{n}, \frac{k}{n}]$, $k = 1,\dots,n$, we may calculate each integral to obtain, for $j = 1$,

$$\int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t)\,\mathrm{d}t = -\frac{(X^*_1)^2}{4n} + \frac{X^*_n(1-X^*_n)}{4n} - \sum_{k=1}^{n-1}\frac{(X^*_{k+1}-X^*_k)^2}{8n}$$

and since $X^*_n(1-X^*_n) = (1-X^*_n) - (1-X^*_n)^2$, Lemma 3.1 implies that

$$\mathbb{E}\int_0^1 (G_{n1}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t)\,\mathrm{d}t = \frac{\mathbb{E}\,(1-X^*_n)}{4n} - \sum_{k=1}^{n-1}\frac{\mathbb{E}\,(X^*_{k+1}-X^*_k)^2}{8n} + \mathcal{O}\big(\frac{1}{n^3}\big)$$

and one may conclude with the relation (8). For $j = 2$, we obtain

$$\int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t)\,\mathrm{d}t = -\frac{(X^*_2-X^*_1)X^*_1}{8n} + \frac{X^*_n(X^*_n-X^*_{n-1})}{8n} - \sum_{k=1}^{n-1}\frac{(X^*_{k+1}-X^*_k)^2}{8n}$$

and, since $(X^*_n - X^*_{n-1})X^*_n = -(X^*_n-X^*_{n-1})(1-X^*_n) + (X^*_n - X^*_{n-1})$, Cauchy-Schwarz inequality and Lemma 3.1 imply that

$$\mathbb{E}\int_0^1 (G_{n2}^{-1}(t) - F_n^{-1}(t))F_n^{-1}(t)\,\mathrm{d}t = \frac{\mathbb{E}\,(X^*_n-X^*_{n-1})}{8n} - \sum_{k=1}^{n-1}\frac{\mathbb{E}\,(X^*_{k+1}-X^*_k)^2}{8n} + \mathcal{O}\big(n^{-3}\big)$$

and again the relation (8) gives the result.

3) This is the most technical term to handle. For $j = 1$, we decompose it into

$$\begin{aligned}\mathbb{E}\,(X^*_1)\int_0^{\frac12}(2nt-1)F^{-1}(t)\,\mathrm{d}t + \mathbb{E}\,(1-X^*_n)\int_{1-\frac{2}{n}}^1 (2nt-2n+1)F^{-1}(t)\,\mathrm{d}t \\ + \sum_{k=1}^{n-1}\mathbb{E}\,(X^*_{k+1}-X^*_k)\int_{\frac{k}{n}-\frac{1}{2n}}^{\frac{k}{n}}(nt-k+\frac12)F^{-1}(t)\,\mathrm{d}t \\ + \sum_{k=1}^{n-1}\mathbb{E}\,(X^*_{k+1}-X^*_k)\int_{\frac{k}{n}}^{\frac{k}{n}+\frac{1}{2n}}(nt-k-\frac12)F^{-1}(t)\,\mathrm{d}t. \quad (9)\end{aligned}$$

We introduce $K_0(t)$ and $K_1(t)$ the primitives of $F^{-1}(t)$ and $tF^{-1}(t)$ and we use Taylor expansions with integral remainder with $K_1''(t) = F^{-1}(t) + \frac{t}{f(F^{-1}(t))}$, $K_1^{(3)}(t) = \frac{2}{f(F^{-1}(t))} - \frac{tf'(F^{-1}(t))}{f^3(F^{-1}(t))}$; and $K_0''(t) = \frac{1}{f(F^{-1}(t))}$, $K_0^{(3)}(t) = -\frac{f'(F^{-1}(t))}{f^3(F^{-1}(t))}$.

By integrating by parts, we arrive at

$$\int_0^{\frac{1}{2n}} (2nt-1)F^{-1}(t)\,\mathrm{d}t = 2n\int_0^{\frac{1}{2n}} \frac{(\frac{1}{2n}-t)^2}{f(F^{-1}(t))}\,\mathrm{d}t - \frac{1}{8n^2f(0)} + \mathcal{O}(n^{-3})$$
$$= \frac{1}{12n^2f(0)} - \frac{1}{8n^2f(0)} + \mathcal{O}(n^{-3}) = -\frac{1}{24n^2f(0)} + \mathcal{O}(n^{-3})$$

so that Lemma 3.1-(a) gives that the first term of (9) is a $\mathcal{O}(n^{-3})$. We use the same methodology for the second term to obtain that

$$\int_{1-\frac{1}{2n}}^1 (2nt-2n+1)F^{-1}(t)\,\mathrm{d}t = \frac{F^{-1}(1-\frac{1}{2n})}{4n} + \frac{1}{12n^2f(F^{-1}(1-\frac{1}{2n}))} + \mathcal{O}(n^{-3})$$
$$= \frac{1}{4n} - \frac{1}{24n^2f(1)} + \mathcal{O}(n^{-3}).$$

We may deduce that the second term of (9) is equal to $\frac{\mathbb{E}\,(1-X_n^*)}{4n} + \mathcal{O}(n^{-3})$ with the help of Lemma 3.1-(d). We use again Taylor expansions with integral remainder together with integration by parts for the two terms depending on k. This allows to get, uniformly in k, that

$$\int_{\frac{k}{n}-\frac{1}{2n}}^{\frac{k}{n}} (nt-k+\frac{1}{2})F^{-1}(t)\,\mathrm{d}t = \frac{F^{-1}(\frac{k}{n}-\frac{1}{2n})}{8n} + \frac{1}{24n^2f(F^{-1}(\frac{k}{n}-\frac{1}{2n}))} + \mathcal{O}(n^{-3})$$

and

$$\int_{\frac{k}{n}}^{\frac{k}{n}+\frac{1}{2n}} (nt-k-\frac{1}{2})F^{-1}(t)\,\mathrm{d}t = -\frac{F^{-1}(\frac{k}{n})}{8n} - \frac{1}{48n^2f(F^{-1}(\frac{k}{n}))} + \mathcal{O}(n^{-3}).$$

By this way,

$$\sum_{k=1}^{n-1} \mathbb{E}\,(X_{k+1}^*-X_k^*)\Big(\int_{\frac{k}{n}-\frac{1}{2n}}^{\frac{k}{n}} (nt-k+\frac{1}{2})F^{-1}(t)\,\mathrm{d}t + \int_{\frac{k}{n}}^{\frac{k}{n}+\frac{1}{2n}} (nt-k-\frac{1}{2})F^{-1}(t)\,\mathrm{d}t\Big)$$
$$= -\frac{1}{24n^2}\mathbb{E}\sum_{k=1}^{n-1} \frac{(X_{k+1}^*-X_k^*)}{f(F^{-1}(\frac{k}{n}))} + \mathcal{O}(n^{-3}).$$

The last task is now to study $\mathbb{E}\,\big(\sum_{k=1}^{n-1} \frac{X_{k+1}^*-X_k^*}{f(F^{-1}(\frac{k}{n}))}\big)$. As $F_n(X_k^*) = \frac{k}{n}$, we have

$$\mathbb{E}\sum_{k=1}^{n-1} \frac{(X_{k+1}^*-X_k^*)}{f(F^{-1}(\frac{k}{n}))} \;=\; \mathbb{E}\,\big(\sum_{k=1}^{n-1} \frac{X_{k+1}^*-X_k^*}{f(X_k^*)}\big) \;+\; \mathbb{E}\,\Big(\sum_{k=1}^{n-1} \frac{(X_{k+1}^*-X_k^*)(f(X_k^*)-f(F^{-1}(\frac{k}{n})))}{f(F^{-1}(\frac{k}{n}))f(X_k^*)}\Big).$$

The first term is evaluated with Lemma 3.2 yielding to

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

Using (7) and the relation $\int_0^1 (1-x)F^{n-1}(x)\,\mathrm{d}x = \frac{(1-\mathbb{E}\,(\sup_{i=1,\dots,n-1} X_i))^2}{2}$ together with Lemma 3.1-(b), we arrive at

$$\mathbb{E}\,\big(\sum_{k=1}^{n-1} \frac{X_{k+1}^*-X_k^*}{f(X_k^*)}\big) = \int_0^1 \frac{1}{f(y)}\,\mathrm{d}y - \frac{1}{2(n+1)f^2(0)} - \frac{1}{2(n+1)f^2(1)}$$
$$- \frac{n(1-\mathbb{E}\,(\sup_{i=1,\dots,n-1} X_i))^2}{2} + \mathcal{O}(n^{-2}).$$

For the second term, using Cauchy-Schwarz inequality implies that it may be bounded by

$$C\big(\sum_{k=1}^{n-1} \mathbb{E}\,(X_{k+1}^*-X_k^*)^2\big)^{\frac{1}{2}}\big(\sum_{k=1}^{n-1} \mathbb{E}\,(F_n^{-1}(\frac{k}{n})-F^{-1}(\frac{k}{n}))^2\big)^{\frac{1}{2}}$$

with C some positive constant. From relation (8), Riemann approximation and Proposition 2.5, this term is of order $\mathcal{O}(n^{-\frac{1}{2}})$. Collecting all the results, the assertion holds for $j = 1$, and is unchanged for $j = 2$, details are omitted. □

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