Lissages polygonaux de la fonction de répartition empirique

Delphine Blanke *(UAPV)* - Denis Bosq *(Paris 6)*

Besançon, 23 novembre 2017

Outline

Introduction

Consider $X_1, \dots, X_n$ i.i.d. and real-valued r.v. with distribution function F and density f. The most classical and natural estimator of F is the edf:

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

$\rightsquigarrow$ unbiased, strongly uniformly consistent, but ... discontinuous

Alternative estimators (Servien, 09):

- Kernel distribution estimator : $K_n(t) = \frac{1}{nh_n}\int_{-\infty}^{t}\sum_{i=1}^{n} k(\frac{x-X_i}{h_n})\,\mathrm{d}x$ with k a classical density kernel
- Other estimators : local smoothing (Lejeune and Sarda, 92), level-crossing (Huang and Brill, 04), splines (Berlinet, 81), ...
- All integrated density estimators ...

Introduction

Consider $X_1, \dots, X_n$ i.i.d. and real-valued r.v. with distribution function F and density f. The most classical and natural estimator of F is the edf:

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

$\rightsquigarrow$ unbiased, strongly uniformly consistent, but ... discontinuous

Alternative estimators (Servien, 09):

- Kernel distribution estimator :
  $K_n(t) = \frac{1}{nh_n} \int_{-\infty}^{t} \sum_{i=1}^{n} k(\frac{x-X_i}{h_n})\,\mathrm{d}x$ with k a classical density kernel
- Other estimators : local smoothing (Lejeune and Sarda, 92), level-crossing (Huang and Brill, 04), splines (Berlinet, 81), ...
- All integrated density estimators ...

Introduction

Consider $X_1, \ldots, X_n$ i.i.d. and real-valued r.v. with distribution function F and density f. The most classical and natural estimator of F is the edf:

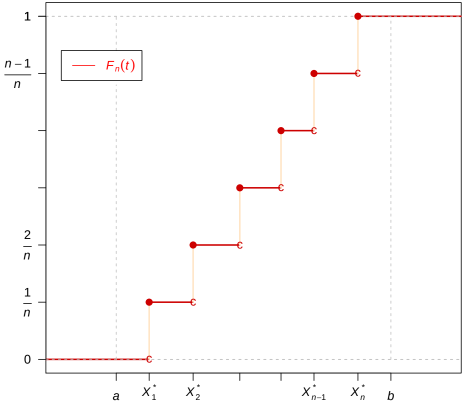
$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{]-\infty,t]}(X_i)$$

$\rightsquigarrow$ unbiased, strongly uniformly consistent, but ... discontinuous

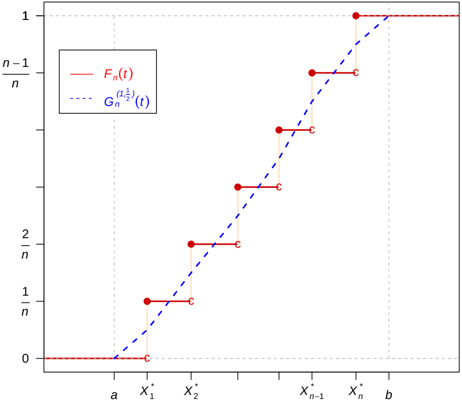
Alternative estimators (Servien, 09):

- Kernel distribution estimator :
  $K_n(t) = \frac{1}{nh_n} \int_{-\infty}^{t} \sum_{i=1}^{n} k(\frac{x-X_i}{h_n}) \, \mathrm{d}x$ with k a classical density kernel
- Other estimators : local smoothing (Lejeune and Sarda, 92), level-crossing (Huang and Brill, 04), splines (Berlinet, 81), ...
- All integrated density estimators ...

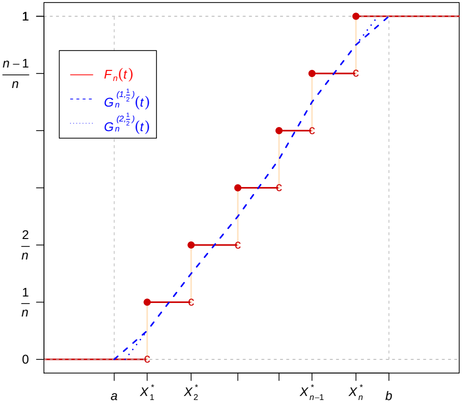
Polygonal estimators



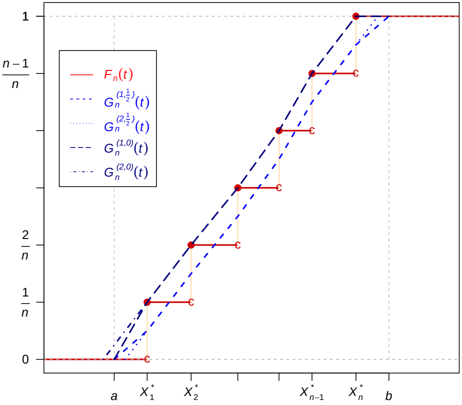
Polygonal estimators



Polygonal estimators

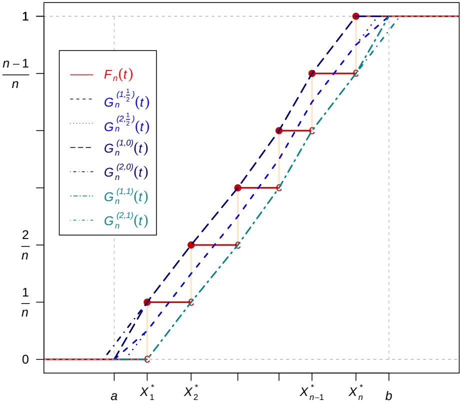


Polygonal estimators



Polygonal estimators

Families $G_n^{(j,p)}$: $j = 1$ known support $[a, b]$, $j = 2$ unknown support and p a known parameter in $[0, 1]$.



Definition (known support $[a, b]$)

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

Definition (unknown support)

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

Definition (known support $[a, b]$)

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

Definition (unknown support)

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

First properties

- $G_n^{(j,p)}$ are continuous cdf and their piecewise derivatives are densities.
- $G_n^{(j,p)}(X_k^*) = \frac{k-p}{n}$ and $G_n^{(j,p)}(X_{k+1}^*) - G_n^{(j,p)}(X_k^*) = \frac{1}{n}$.
- $G_n^{(j,p)}(t) = F_n(t)$ for $t = X_k^* + p(X_{k+1}^* - X_k^*)$.
- $G_n^{(1,p)}(a) = F_n(a) = 0$ and $G_n^{(1,p)}(b) = F_n(b) = 1$.
- $G_n^{(2,p)}(t) \equiv 0$ for $t \leq (2-p)X_1^* - (1-p)X_2^* \leq X_1^*$ and $G_n^{(2,p)}(t) \equiv 1$ for $t \geq (1+p)X_n^* - pX_{n-1}^* \geq X_n^*$ but $\left[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\right] \not\subset [a,b]$.

First properties

- $G_n^{(j,p)}$ are continuous cdf and their piecewise derivatives are densities.
- $G_n^{(j,p)}(X_k^*) = \frac{k-p}{n}$ and $G_n^{(j,p)}(X_{k+1}^*) - G_n^{(j,p)}(X_k^*) = \frac{1}{n}$.
- $G_n^{(j,p)}(t) = F_n(t)$ for $t = X_k^* + p(X_{k+1}^* - X_k^*)$.
- $G_n^{(1,p)}(a) = F_n(a) = 0$ and $G_n^{(1,p)}(b) = F_n(b) = 1$.
- $G_n^{(2,p)}(t) \equiv 0$ for $t \leq (2-p)X_1^* - (1-p)X_2^* \leq X_1^*$ and $G_n^{(2,p)}(t) \equiv 1$ for $t \geq (1+p)X_n^* - pX_{n-1}^* \geq X_n^*$ but $\left[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\right] \not\subset [a, b]$.

First properties

- $G_n^{(j,p)}$ are continuous cdf and their piecewise derivatives are densities.
- $G_n^{(j,p)}(X_k^*) = \frac{k-p}{n}$ and $G_n^{(j,p)}(X_{k+1}^*) - G_n^{(j,p)}(X_k^*) = \frac{1}{n}$.
- $G_n^{(j,p)}(t) = F_n(t)$ for $t = X_k^* + p(X_{k+1}^* - X_k^*)$.
- $G_n^{(1,p)}(a) = F_n(a) = 0$ and $G_n^{(1,p)}(b) = F_n(b) = 1$.
- $G_n^{(2,p)}(t) \equiv 0$ for $t \leq (2-p)X_1^* - (1-p)X_2^* \leq X_1^*$ and $G_n^{(2,p)}(t) \equiv 1$ for $t \geq (1+p)X_n^* - pX_{n-1}^* \geq X_n^*$ but $\left[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\right] \not\subset [a, b]$.

First properties

- $G_n^{(j,p)}$ are continuous cdf and their piecewise derivatives are densities.
- $G_n^{(j,p)}(X_k^*) = \frac{k-p}{n}$ and $G_n^{(j,p)}(X_{k+1}^*) - G_n^{(j,p)}(X_k^*) = \frac{1}{n}$.
- $G_n^{(j,p)}(t) = F_n(t)$ for $t = X_k^* + p(X_{k+1}^* - X_k^*)$.
- $G_n^{(1,p)}(a) = F_n(a) = 0$ and $G_n^{(1,p)}(b) = F_n(b) = 1$.
- $G_n^{(2,p)}(t) \equiv 0$ for $t \leq (2-p)X_1^* - (1-p)X_2^* \leq X_1^*$ and $G_n^{(2,p)}(t) \equiv 1$ for $t \geq (1+p)X_n^* - pX_{n-1}^* \geq X_n^*$ but $\left[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\right] \not\subset [a,b]$.

First properties

- $G_n^{(j,p)}$ are continuous cdf and their piecewise derivatives are densities.
- $G_n^{(j,p)}(X_k^*) = \frac{k-p}{n}$ and $G_n^{(j,p)}(X_{k+1}^*) - G_n^{(j,p)}(X_k^*) = \frac{1}{n}$.
- $G_n^{(j,p)}(t) = F_n(t)$ for $t = X_k^* + p(X_{k+1}^* - X_k^*)$.
- $G_n^{(1,p)}(a) = F_n(a) = 0$ and $G_n^{(1,p)}(b) = F_n(b) = 1$.
- $G_n^{(2,p)}(t) \equiv 0$ for $t \leq (2-p)X_1^* - (1-p)X_2^* \leq X_1^*$ and $G_n^{(2,p)}(t) \equiv 1$ for $t \geq (1+p)X_n^* - pX_{n-1}^* \geq X_n^*$ but $\left[(2-p)X_1^* - (1-p)X_2^*, (1+p)X_n^* - pX_{n-1}^*\right] \not\subset [a, b]$.

Only one reference (?) about polygonal estimators (Read, 72).

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

It is shown that, for n sufficiently large, the expected squared error of $G_n^{(1,0)}$ is no larger than F_n. Also, a variant of $G_n^{(1,0)}$ dominates $\frac{nF_n+1}{n+2}$ in terms of integrated risk but the result is not proven.

In all the following and, without loss of generality, $[a, b] \equiv [0, 1]$

Lemma

For the families of estimators $G_n^{(1,p)}$, we get

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

Lemma

For the families of estimators $G_n^{(2,p)}$, we get

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

Lemma

For the families of estimators $G_n^{(2,p)}$, we get

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

$\rightsquigarrow \left\| F_n - G_n^{(j,p)} \right\|_\infty = \max\left(\frac{p}{n}, \frac{1-p}{n}\right),\ 0 \le p \le 1,\ j = 1, 2.$

Exponential inequalities

Proposition

If F is continuous and increasing over [0,1], 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. The result is derived from

$$\left\| G_n^{(j,p)} - F \right\|_\infty \leq \frac{\max(p, 1-p)}{n} + \|F_n - F\|_\infty$$

and the exponential inequality (Massart, 90)

$$\mathbb{P}\big(\|F_n - F\|_\infty \geq \varepsilon \big) \leq 2 \exp\big(-2n\varepsilon^2 \big)$$

with the choice $\varepsilon > \frac{\max(p,1-p)}{n(1-a_0)}$, $n \geq 1$, $0 < a_0 < 1$.

Proposition

Under same conditions as before, we get

(a) for $p = \frac{1}{2}$, $j = 1, 2$:

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

(b) and more generally,

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

$0 < \varepsilon < \max\left(\frac{p}{2n}, \frac{1-p}{2n}\right),\ n \geq 1.$

Proof. The result is derived again from the choice of ε and

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

Proposition

Under same conditions as before, we get

(a) for $p = \frac{1}{2}$, $j = 1, 2$:

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

(b) and more generally,

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

$0 < \varepsilon < \max\left(\frac{p}{2n}, \frac{1-p}{2n}\right),\ n \geq 1.$

Proof. The result is derived again from the choice of ε and

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

Study of the MISE

The MISE is calculated from $\mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(j,p)}(t) - F(t)\right)^2 \mathrm{d}t$

$$= \mathrm{E} \int_{-\infty}^{\infty} \left(F_n(t) - F(t)\right)^2 \mathrm{d}t + \mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(j,p)}(t) - F_n(t)\right)^2 \mathrm{d}t$$

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

Lemma

(a) *For* $m \in \mathbb{N}^*$, $\int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right)^m \mathrm{d}t$

$$= \frac{\left((1-p)^m - (-1)^m p^m\right)\left(pX_1^* + (1-p)X_n^*\right) + (-1)^m p^m}{(m+1)n^m}.$$

Study of the MISE

The MISE is calculated from $\mathrm{E}\int_{-\infty}^{\infty}\left(G_n^{(j,p)}(t)-F(t)\right)^2\mathrm{d}t$

$$= \mathrm{E}\int_{-\infty}^{\infty}\left(F_n(t)-F(t)\right)^2\mathrm{d}t + \mathrm{E}\int_{-\infty}^{\infty}\left(G_n^{(j,p)}(t)-F_n(t)\right)^2\mathrm{d}t$$

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

Lemma

(a) *For* $m\in\mathbb{N}^*$, $\int_0^1\left(G_n^{(1,p)}(t)-F_n(t)\right)^m\mathrm{d}t$

$$=\frac{\left((1-p)^m-(-1)^m p^m\right)\left(pX_1^*+(1-p)X_n^*\right)+(-1)^m p^m}{(m+1)n^m}.$$

Study of the MISE

The MISE is calculated from $\mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(j,p)}(t) - F(t)\right)^2 \mathrm{d}t$

$$= \mathrm{E} \int_{-\infty}^{\infty} \left(F_n(t) - F(t)\right)^2 \mathrm{d}t + \mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(j,p)}(t) - F_n(t)\right)^2 \mathrm{d}t$$

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

Lemma

(a) *For* $m \in \mathbb{N}^*$, $\int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right)^m \mathrm{d}t$

$$= \frac{\left((1-p)^m - (-1)^m p^m\right)\left(pX_1^* + (1-p)X_n^*\right) + (-1)^m p^m}{(m+1)n^m}.$$

$\rightsquigarrow$ *for* m *even and* $p = \frac{1}{2}$, *the result is* $\frac{(2n)^{-m}}{m+1}$!

Study of the MISE

The MISE is calculated from $\mathrm{E}\int_{-\infty}^{\infty}\left(G_n^{(j,p)}(t)-F(t)\right)^2\mathrm{d}t$

$$= \mathrm{E}\int_{-\infty}^{\infty}\left(F_n(t)-F(t)\right)^2\mathrm{d}t + \mathrm{E}\int_{-\infty}^{\infty}\left(G_n^{(j,p)}(t)-F_n(t)\right)^2\mathrm{d}t$$

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

Lemma

(a) *For* $m\in\mathbb{N}^*$, $\int_0^1\left(G_n^{(1,p)}(t)-F_n(t)\right)^m\mathrm{d}t$

$$=\frac{\left((1-p)^m-(-1)^m p^m\right)\left(pX_1^*+(1-p)X_n^*\right)+(-1)^m p^m}{(m+1)n^m}.$$

(b) *A similar but more complicated expression is obtained for* $j=2$ *involving* X_2^* *and* X_{n-1}^* *too.*

Assumption (A1)

(i) *F admits the density f supposed to be compactly supported on [0,1]*

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

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

Lemma

If the conditions A1-(i)(ii) hold then, for all integers $r \geq 0$ and $m \geq 1$, not depending on n, we get

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

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

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

$\mathrm{E}\left(X_2^* - X_1^*\right)^m = \mathcal{O}\left(\frac{1}{n^m}\right),$

(d) $\mathrm{E}\left(X_n^* - X_{n-1}^*\right) = \frac{e_1}{n} + \mathcal{O}\left(\frac{1}{n^2}\right)$, $e_1 > 0$, *and*

$\mathrm{E}\left(X_n^* - X_{n-1}^*\right)^m = \mathcal{O}\left(\frac{1}{n^m}\right).$

Proposition

Under the conditions A1-(i)(ii), we have for all $p \in [0,1]$

(a) $\mathrm{E} \int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right)^2 \mathrm{d}t$

$$= \frac{(1-2p)\left(p\mathrm{E}\left(X_1^*\right) - (1-p)\mathrm{E}\left(1-X_n^*\right)\right) + 1 - 3p + 3p^2}{3n^2}$$

$$= \frac{1-3p+3p^2}{3n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big);$$

(b) $\mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(2,p)}(t) - F_n(t)\right)^2 \mathrm{d}t$

$$= \frac{p^3\mathrm{E}\left(X_n^* - X_{n-1}^*\right) + \left((1-p)^3 + p^3\right)}{3n^2}$$

$$+ \frac{(1-p)^3\mathrm{E}\left(X_2^* - X_1^*\right) + \mathrm{E}\left(X_n^* - X_1^*\right)\left((1-p)^3 + p^3\right)}{3n^2}$$

$$= \frac{1-3p+3p^2}{3n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big).$$

Proposition

Under the conditions A1-(i)(ii), we have for all $p \in [0,1]$

(a) $\mathrm{E} \int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right)^2 \mathrm{d}t$

$$= \frac{(1-2p)\big(p\mathrm{E}\,(X_1^*) - (1-p)\mathrm{E}\,(1-X_n^*)\big) + 1 - 3p + 3p^2}{3n^2}$$

$$= \frac{1-3p+3p^2}{3n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big);$$

(b) $\mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(2,p)}(t) - F_n(t)\right)^2 \mathrm{d}t$

$$= \frac{p^3\mathrm{E}\,(X_n^* - X_{n-1}^*) + \left((1-p)^3 + p^3\right)}{3n^2}$$

$$+ \frac{(1-p)^3\mathrm{E}\,(X_2^* - X_1^*) + \mathrm{E}\,(X_n^* - X_1^*)\left((1-p)^3 + p^3\right)}{3n^2}$$

$$= \frac{1-3p+3p^2}{3n^2} + \mathcal{O}\Big(\frac{1}{n^3}\Big).$$

Now, the most difficult task is the double product. We get

Proposition

Under Assumption A1, we get for $j = 1, 2$ and $p \in [0, 1]$:

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

Sketch of Proof 1/3, $j = 1$.

$I_1^{(1,p)} = 2\,\mathrm{E} \int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right) F_n(t)\,\mathrm{d}t$ is quite easy to derive as F_n is piecewise constant. We obtain

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

$$= \frac{(1-2p)(1-\mathrm{E}(X_1))}{n} - \frac{(1-p)b_1}{n^2} + \mathcal{O}(\frac{1}{n^3})$$

Now, the most difficult task is the double product. We get

Proposition

Under Assumption A1, we get for $j = 1, 2$ and $p \in [0, 1]$:

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

Sketch of Proof 1/3, $j = 1$.

$I_1^{(1,p)} = 2\,\mathrm{E} \int_0^1 \left(G_n^{(1,p)}(t) - F_n(t)\right) F_n(t)\,\mathrm{d}t$ is quite easy to derive as F_n is piecewise constant. We obtain

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

$$= \frac{(1-2p)(1-\mathrm{E}(X_1))}{n} - \frac{(1-p)b_1}{n^2} + \mathcal{O}(\frac{1}{n^3})$$

Now, the most difficult task is the double product. We get

Proposition

Under Assumption A1, we get for $j = 1, 2$ and $p \in [0, 1]$:

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

Sketch of Proof 1/3, $j = 1$.

$I_1^{(1,p)} = 2\,\mathrm{E} \int_0^1 \big(G_n^{(1,p)}(t) - F_n(t)\big) F_n(t)\,\mathrm{d}t$ is quite easy to derive as F_n is piecewise constant. We obtain

$$I_1^{(1,p)} = -\mathrm{E}\Big(\sum_{k=1}^{n-1} \frac{(2p-1)k(X_{k+1}^* - X_k^*)}{n^2}\Big) - \frac{p(1 - \mathrm{E}(X_n^*))}{n}$$
$$= \frac{(1-2p)(1-\mathrm{E}(X_1))}{n} - \frac{(1-p)b_1}{n^2} + \mathcal{O}(\frac{1}{n^3})$$

Sketch of Proof 2/3, $j = 1$.

$I_2^{(1,p)} = 2\,\mathrm{E}\,\int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t$ is the most technical term.

$$I_2^{(1,p)} = \mathrm{E}\Bigg(\frac{-2(1-p)(X_1^*)^2\big(f(0) + X_1^* R_{1,0}\big)}{3n} + \frac{p(1-X_n^*)\big(3F(X_n^*) + (1-X_n^*)(f(X_n^*) + (1-X_n^*)R_{1,n})\big)}{3n} + \frac{(2p-1)}{n}\sum_{k=1}^{n-1}(X_{k+1}^* - X_k^*)F(X_k^*) + \frac{(3p-2)}{3n}\sum_{k=1}^{n-1}(X_{k+1}^* - X_k^*)^2\big(f(X_k^*) + (X_{k+1}^* - X_k^*)R_{1,k}\big)\Bigg)$$

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

Sketch of Proof 2/3, $j = 1$.

$I_2^{(1,p)} = 2\,\mathrm{E} \int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big) F(t)\,\mathrm{d}t$ is the most technical term.

$$I_2^{(1,p)} = \mathrm{E}\Bigg(\frac{-2(1-p)(X_1^*)^2\big(f(0) + X_1^* R_{1,0}\big)}{3n} + \frac{p(1-X_n^*)\big(3F(X_n^*) + (1-X_n^*)(f(X_n^*) + (1-X_n^*)R_{1,n})\big)}{3n} + \frac{(2p-1)}{n}\sum_{k=1}^{n-1}(X_{k+1}^* - X_k^*)F(X_k^*) + \frac{(3p-2)}{3n}\sum_{k=1}^{n-1}(X_{k+1}^* - X_k^*)^2\big(f(X_k^*) + (X_{k+1}^* - X_k^*)R_{1,k}\big)\Bigg)$$

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

Sketch of Proof 2/3, $j = 1$.

$I_2^{(1,p)} = 2\,\mathrm{E} \int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t$ is the most technical term.

$$I_2^{(1,p)} = \mathrm{E}\Bigg(\frac{-2(1-p)(X_1^*)^2\big(f(0) + X_1^* R_{1,0}\big)}{3n} + \frac{p(1-X_n^*)\big(3F(X_n^*) + (1-X_n^*)(f(X_n^*) + (1-X_n^*)R_{1,n})\big)}{3n} + \frac{(2p-1)}{n}\sum_{k=1}^{n-1}(X_{k+1}^* - X_k^*)F(X_k^*) + \frac{(3p-2)}{3n}\sum_{k=1}^{n-1}(X_{k+1}^* - X_k^*)^2\big(f(X_k^*) + (X_{k+1}^* - X_k^*)R_{1,k}\big)\Bigg)$$

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

Sketch of Proof 2/3, $j = 1$.

$I_2^{(1,p)} = 2\,\mathrm{E}\int_0^1 \big(F_n(t) - G_n^{(1,p)}(t)\big)F(t)\,\mathrm{d}t$ is the most technical term.

$$I_2^{(1,p)} = \mathrm{E}\Bigg(\frac{-2(1-p)(X_1^*)^2\big(f(0)+X_1^*R_{1,0}\big)}{3n} + \frac{p(1-X_n^*)\big(3F(X_n^*)+(1-X_n^*)(f(X_n^*)+(1-X_n^*)R_{1,n})\big)}{3n} + \frac{(2p-1)}{n}\sum_{k=1}^{n-1}(X_{k+1}^*-X_k^*)F(X_k^*) + \frac{(3p-2)}{3n}\sum_{k=1}^{n-1}(X_{k+1}^*-X_k^*)^2\big(f(X_k^*)+(X_{k+1}^*-X_k^*)R_{1,k}\big)\Bigg)$$

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

Sketch of Proof 3/3, $j = 1$.

From the binomial theorem and the joint density of (X_k^*, X_{k+1}^*):

$f_{(X_k^*, X_{k+1}^*)}(x, y) = \frac{n! F^{k-1}(x) f(x) f(y) (1-F(y))^{n-k-1}}{(k-1)!(n-k-1)!}$ with $y > x$, we get

Proposition

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

$$\sum_{k=1}^{n-1} \mathrm{E}\left(h(X_k^*, X_{k+1}^*)\right)$$

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

and after some calculations (...), one arrives at

$$I_2^{(1,p)} = -\frac{(1-2p)(1-\mathrm{E}(X_1))}{n} + \frac{b_1(1-p)}{n^2} - \frac{1}{3n^2} + \mathcal{O}(\frac{1}{n^3}).$$

Sketch of Proof 3/3, $j = 1$.

From the binomial theorem and the joint density of (X_k^*, X_{k+1}^*):
$f_{(X_k^*, X_{k+1}^*)}(x,y) = \frac{n! F^{k-1}(x) f(x) f(y) (1-F(y))^{n-k-1}}{(k-1)!(n-k-1)!}$ with $y > x$, we get

Proposition

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

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

and after some calculations (...), one arrives at

$$I_2^{(1,p)} = -\frac{(1-2p)(1-\mathrm{E}(X_1))}{n} + \frac{b_1(1-p)}{n^2} - \frac{1}{3n^2} + \mathcal{O}(\frac{1}{n^3}).$$

Theorem

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

$$\mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(j,p)}(t) - F(t)\right)^2 \mathrm{d}t$$

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

- $G_n^{(1,p)}$ and $G_n^{(2,p)}$ are asymptotically equivalent.
- For all $p \in]0,1[$, the families $G_n^{(j,p)}$, $j = 1, 2$ are more efficient than F_n.
- Choices $p = 0$ or $p = 1$ are more problematic since the term $\frac{p(1-p)}{n^2}$ vanishes in these cases.
- The better efficiency is achieved for $p = \frac{1}{2}$ where $\frac{p(1-p)}{n^2} = \frac{1}{4n^2}$.

Theorem

Under Assumption 1, we get for $j = 1, 2$ and all $p \in [0, 1]$:
$\mathrm{E} \int_{-\infty}^{\infty} \left(G_n^{(j,p)}(t) - F(t)\right)^2 \mathrm{d}t$

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

- $G_n^{(1,p)}$ and $G_n^{(2,p)}$ are asymptotically equivalent.
- For all $p \in]0, 1[$, the families $G_n^{(j,p)}$, $j = 1, 2$ are more efficient than F_n.
- Choices $p = 0$ or $p = 1$ are more problematic since the term $\frac{p(1-p)}{n^2}$ vanishes in these cases.
- The better efficiency is achieved for $p = \frac{1}{2}$ where $\frac{p(1-p)}{n^2} = \frac{1}{4n^2}$.

The nonparametric kernel distribution estimator is defined as follows

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

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: see Swanepoel and Van Graan, 05 or Servien, 09 for a rich literature review.

- Weighted MISE: $\mathrm{E}\int_{-\infty}^{\infty}\big(K_n(t)-F(t)\big)^2 f(t)\,\mathrm{d}t$ established by Swanepoel, 88 with optimal choice of k.
- Unweighted MISE derived in Jones, 90 when F has two continuous derivatives f and f':

$$\mathrm{E}\int_{-\infty}^{\infty}\big(K_n(t)-F(t)\big)^2\,\mathrm{d}t$$

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

For f only Lipschitz and compactly supported on [0,1], we obtain
$\mathrm{E}\int_0^1 \big(K_n(t) - F(t)\big)^2\,\mathrm{d}t =$

$$\frac{\int_0^1 F(t)(1-F(t))\,\mathrm{d}t}{n} - \frac{2h_n}{n}\int_{-\infty}^{\infty} tk(t)L(t)\,\mathrm{d}t + \mathcal{O}\big(h_n^4\big) + o\big(\frac{h_n}{n}\big).$$

- Similar expression as for $G_n^{(j,p)}$ with presence of the MISE of F_n
- Bandwidth to calibrate: h_n of order $n^{-\frac{1}{3}}$ gives a $\mathcal{O}(n^{-4/3})$ while the improvement is only $\mathcal{O}(n^{-2})$ for $G_n^{(j,p)}$, $p \in]0,1[$.
- Practical choice of h_n?
  - Sarda, 93: leave-one-out cross-validation method
  - Bowman, Hall, Prvan 98: modified cross-validation method
  - Altman and Leger, 95 or Polansky and Baker, 00: plug-in bandwidth choice

For f only Lipschitz and compactly supported on [0,1], we obtain
$\mathrm{E}\int_0^1 \left(K_n(t)-F(t)\right)^2 \mathrm{d}t =$

$$\frac{\int_0^1 F(t)(1-F(t))\,\mathrm{d}t}{n} - \frac{2h_n}{n}\int_{-\infty}^{\infty} tk(t)L(t)\,\mathrm{d}t + \mathcal{O}\left(h_n^4\right) + o\left(\frac{h_n}{n}\right).$$

- Similar expression as for $G_n^{(j,p)}$ with presence of the MISE of F_n
- Bandwidth to calibrate: h_n of order $n^{-\frac{1}{3}}$ gives a $\mathcal{O}(n^{-4/3})$ while the improvement is only $\mathcal{O}(n^{-2})$ for $G_n^{(j,p)}$, $p \in]0,1[$.
- Practical choice of h_n?
  - Sarda, 93: leave-one-out cross-validation method
  - Bowman, Hall, Prvan 98: modified cross-validation method
  - Altman and Leger, 95 or Polansky and Baker, 00: plug-in bandwidth choice

For f only Lipschitz and compactly supported on [0,1], we obtain
$\mathrm{E}\int_0^1 \left(K_n(t)-F(t)\right)^2 \mathrm{d}t =$

$$\frac{\int_0^1 F(t)(1-F(t))\,\mathrm{d}t}{n} - \frac{2h_n}{n}\int_{-\infty}^{\infty} tk(t)L(t)\,\mathrm{d}t + \mathcal{O}\left(h_n^4\right) + o\left(\frac{h_n}{n}\right).$$

- Similar expression as for $G_n^{(j,p)}$ with presence of the MISE of F_n
- Bandwidth to calibrate: h_n of order $n^{-\frac{1}{3}}$ gives a $\mathcal{O}(n^{-4/3})$ while the improvement is only $\mathcal{O}(n^{-2})$ for $G_n^{(j,p)}$, $p \in]0,1[$.
- Practical choice of h_n?
  - Sarda, 93: leave-one-out cross-validation method
  - Bowman, Hall, Prvan 98: modified cross-validation method
  - Altman and Leger, 95 or Polansky and Baker, 00: plug-in bandwidth choice

For f only Lipschitz and compactly supported on [0,1], we obtain
$\mathrm{E}\int_0^1 \big(K_n(t) - F(t)\big)^2 \,\mathrm{d}t =$

$$\frac{\int_0^1 F(t)(1-F(t))\,\mathrm{d}t}{n} - \frac{2h_n}{n}\int_{-\infty}^{\infty} tk(t)L(t)\,\mathrm{d}t + \mathcal{O}\big(h_n^4\big) + o\big(\frac{h_n}{n}\big).$$

- Similar expression as for $G_n^{(j,p)}$ with presence of the MISE of F_n
- Bandwidth to calibrate: h_n of order $n^{-\frac{1}{3}}$ gives a $\mathcal{O}(n^{-4/3})$ while the improvement is only $\mathcal{O}(n^{-2})$ for $G_n^{(j,p)}$, $p \in]0,1[$.
- Practical choice of h_n?
  - Sarda, 93: leave-one-out cross-validation method
  - Bowman, Hall, Prvan 98: modified cross-validation method
  - Altman and Leger, 95 or Polansky and Baker, 00: plug-in bandwidth choice

Multi-stage procedure of Polansky and Baker, 00:

- h_n minimizing the MISE given by

$$h_{\mathrm{opt}} = \left(\frac{2\int tk(t)L(t)\,\mathrm{d}t}{n(\int t^2 k(t)\,\mathrm{d}t)^2 \int (f'(t))^2\,\mathrm{d}t}\right)^{\frac{1}{3}}$$

- Nonparametric kernel estimation of $\int (f'(t))^2\,\mathrm{d}t$ involves a bandwidth h_{1n} with $h_{1,\mathrm{opt}}$ depending on $\int (f^{(2)}(t))^2\,\mathrm{d}t$
- Nonparametric kernel estimation of $\int (f^{(2)}(t))^2\,\mathrm{d}t$ involves a bandwidth h_{2n} with $h_{2,\mathrm{opt}}$ depending on $\int (f^{(3)}(t))^2\,\mathrm{d}t$ and so on ...
- For a two-stage procedure, $\int (f^{(3)}(t))^2\,\mathrm{d}t$ estimated with a reference distribution, namely normal with variance σ^2, and $\widehat{\sigma} = \min(S_n, \frac{\widehat{q}_{0.75}-\widehat{q}_{0.25}}{1.349})$.

Multi-stage procedure of Polansky and Baker, 00:

- h_n minimizing the MISE given by
$$h_{\text{opt}} = \left(\frac{2 \int t k(t) L(t)\,\mathrm{d}t}{n(\int t^2 k(t)\,\mathrm{d}t)^2 \int (f'(t))^2\,\mathrm{d}t} \right)^{\frac{1}{3}}$$
- Nonparametric kernel estimation of $\int (f'(t))^2\,\mathrm{d}t$ involves a bandwidth h_{1n} with $h_{1,\text{opt}}$ depending on $\int (f^{(2)}(t))^2\,\mathrm{d}t$
- Nonparametric kernel estimation of $\int (f^{(2)}(t))^2\,\mathrm{d}t$ involves a bandwidth h_{2n} with $h_{2,\text{opt}}$ depending on $\int (f^{(3)}(t))^2\,\mathrm{d}t$ and so on ...
- For a two-stage procedure, $\int (f^{(3)}(t))^2\,\mathrm{d}t$ estimated with a reference distribution, namely normal with variance σ^2, and $\widehat{\sigma} = \min(S_n, \frac{\widehat{q}_{0.75} - \widehat{q}_{0.25}}{1.349})$.

Multi-stage procedure of Polansky and Baker, 00:

- h_n minimizing the MISE given by

$$h_{\text{opt}} = \left(\frac{2 \int t k(t) L(t) \, \mathrm{d}t}{n (\int t^2 k(t) \, \mathrm{d}t)^2 \int (f'(t))^2 \, \mathrm{d}t} \right)^{\frac{1}{3}}$$

- Nonparametric kernel estimation of $\int (f'(t))^2 \, \mathrm{d}t$ involves a bandwidth h_{1n} with $h_{1,\text{opt}}$ depending on $\int (f^{(2)}(t))^2 \, \mathrm{d}t$
- Nonparametric kernel estimation of $\int (f^{(2)}(t))^2 \, \mathrm{d}t$ involves a bandwidth h_{2n} with $h_{2,\text{opt}}$ depending on $\int (f^{(3)}(t))^2 \, \mathrm{d}t$ and so on ...
- For a two-stage procedure, $\int (f^{(3)}(t))^2 \, \mathrm{d}t$ estimated with a reference distribution, namely normal with variance σ^2, and $\widehat{\sigma} = \min(S_n, \frac{\widehat{q}_{0.75} - \widehat{q}_{0.25}}{1.349})$.

Multi-stage procedure of Polansky and Baker, 00:

- h_n minimizing the MISE given by

$$h_{\text{opt}} = \left(\frac{2 \int t k(t) L(t) \, \mathrm{d}t}{n (\int t^2 k(t) \, \mathrm{d}t)^2 \int (f'(t))^2 \, \mathrm{d}t} \right)^{\frac{1}{3}}$$

- Nonparametric kernel estimation of $\int (f'(t))^2 \, \mathrm{d}t$ involves a bandwidth h_{1n} with $h_{1,\text{opt}}$ depending on $\int (f^{(2)}(t))^2 \, \mathrm{d}t$
- Nonparametric kernel estimation of $\int (f^{(2)}(t))^2 \, \mathrm{d}t$ involves a bandwidth h_{2n} with $h_{2,\text{opt}}$ depending on $\int (f^{(3)}(t))^2 \, \mathrm{d}t$ and so on ...
- For a two-stage procedure, $\int (f^{(3)}(t))^2 \, \mathrm{d}t$ estimated with a reference distribution, namely normal with variance σ^2, and $\widehat{\sigma} = \min(S_n, \frac{\widehat{q}_{0.75} - \widehat{q}_{0.25}}{1.349})$.

- Numerical computation of F_n, K_n (Gaussian kernel k, h_n chosen with 2-stage Polansky and Baker procedure), $G_n^{(j,p)}$, $j = 1, 2$ with $p = \frac{1}{2}$ but also $p = 0$ or 1
- Set of 15 Gaussian mixtures defined in Marron and Wand, 92 + 1 additional from Janssen, Marron, Veraverbeke, and Sarle, 95
  - of easy implementation,
  - describing a broad class of potential problems (skewness, multimodality, and heavy kurtosis)
  - parameters chosen such that $\min_{\ell=1,\dots,16}(\mu_\ell - 3\sigma_\ell) = -3$ and $\max_{\ell=1,\dots,16}(\mu_\ell + 3\sigma_\ell) = 3$
- $N = 500$ samples of sizes $n = 20$, 50 and 100 are generated and a Monte Carlo approximation is operated for each sample to estimate the ISE
- $\widehat{\text{MISE}}$, is obtained by averaging the results over the N replicates

- Numerical computation of F_n, K_n (Gaussian kernel k, h_n chosen with 2-stage Polansky and Baker procedure), $G_n^{(j,p)}$, $j = 1, 2$ with $p = \frac{1}{2}$ but also $p = 0$ or 1
- Set of 15 Gaussian mixtures defined in Marron and Wand, 92 + 1 additional from Janssen, Marron, Veraverbeke, and Sarle, 95
  - of easy implementation,
  - describing a broad class of potential problems (skewness, multimodality, and heavy kurtosis)
  - parameters chosen such that $\min_{\ell=1,\dots,16}(\mu_\ell - 3\sigma_\ell) = -3$ and $\max_{\ell=1,\dots,16}(\mu_\ell + 3\sigma_\ell) = 3$
- $N = 500$ samples of sizes $n = 20$, 50 and 100 are generated and a Monte Carlo approximation is operated for each sample to estimate the ISE
- $\widehat{\text{MISE}}$, is obtained by averaging the results over the N replicates

- Numerical computation of F_n, K_n (Gaussian kernel k, h_n chosen with 2-stage Polansky and Baker procedure), $G_n^{(j,p)}$, $j = 1, 2$ with $p = \frac{1}{2}$ but also $p = 0$ or 1
- Set of 15 Gaussian mixtures defined in Marron and Wand, 92 + 1 additional from Janssen, Marron, Veraverbeke, and Sarle, 95
  - of easy implementation,
  - describing a broad class of potential problems (skewness, multimodality, and heavy kurtosis)
  - parameters chosen such that $\min_{\ell=1,\ldots,16}(\mu_\ell - 3\sigma_\ell) = -3$ and $\max_{\ell=1,\ldots,16}(\mu_\ell + 3\sigma_\ell) = 3$
- $N = 500$ samples of sizes $n = 20$, 50 and 100 are generated and a Monte Carlo approximation is operated for each sample to estimate the ISE
- $\widehat{\text{MISE}}$, is obtained by averaging the results over the N replicates

- Numerical computation of F_n, K_n (Gaussian kernel k, h_n chosen with 2-stage Polansky and Baker procedure), $G_n^{(j,p)}$, $j = 1, 2$ with $p = \frac{1}{2}$ but also $p = 0$ or 1
- Set of 15 Gaussian mixtures defined in Marron and Wand, 92 + 1 additional from Janssen, Marron, Veraverbeke, and Sarle, 95
    - of easy implementation,
    - describing a broad class of potential problems (skewness, multimodality, and heavy kurtosis)
    - parameters chosen such that $\min_{\ell=1,\dots,16}(\mu_\ell - 3\sigma_\ell) = -3$ and $\max_{\ell=1,\dots,16}(\mu_\ell + 3\sigma_\ell) = 3$
- $N = 500$ samples of sizes $n = 20$, 50 and 100 are generated and a Monte Carlo approximation is operated for each sample to estimate the ISE
- $\widehat{\text{MISE}}$, is obtained by averaging the results over the N replicates

Results

- For all tested distributions and all n in $\{20, 50, 100\}$, $G_n^{(1,\frac{1}{2})}$ and $G_n^{(2,\frac{1}{2})}$ outperform F_n
- $G_n^{(1,\frac{1}{2})}$ with $a = -3$ and $b = 3$ always slightly better than $G_n^{(2,\frac{1}{2})}$
- $G_n^{(j,0)}$ and $G_n^{(j,1)}$, $j = 1, 2$ have irregular behaviour: better or worse than F_n depending on n and the simulated distribution. Nevertheless, their estimated MISE are always greater than $G_n^{(j,\frac{1}{2})}$
- $G_n^{(1,\frac{1}{2})}$ outperforms both K_n and F_n for 6/16 tested distributions (from only $n = 50$ for 2 of them)

Results

- For all tested distributions and all n in $\{20, 50, 100\}$, $G_n^{(1,\frac{1}{2})}$ and $G_n^{(2,\frac{1}{2})}$ outperform F_n
- $G_n^{(1,\frac{1}{2})}$ with $a = -3$ and $b = 3$ always slightly better than $G_n^{(2,\frac{1}{2})}$
- $G_n^{(j,0)}$ and $G_n^{(j,1)}$, $j = 1, 2$ have irregular behaviour: better or worse than F_n depending on n and the simulated distribution. Nevertheless, their estimated MISE are always greater than $G_n^{(j,\frac{1}{2})}$
- $G_n^{(1,\frac{1}{2})}$ outperforms both K_n and F_n for 6/16 tested distributions (from only $n = 50$ for 2 of them)

Results

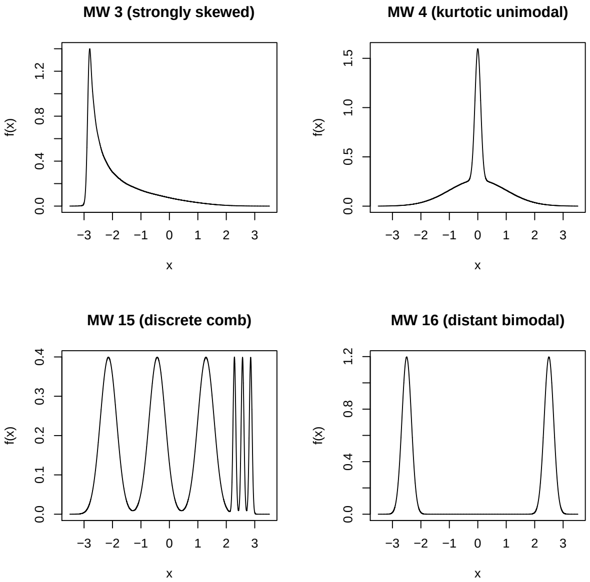
- For all tested distributions and all n in $\{20, 50, 100\}$, $G_n^{(1,\frac{1}{2})}$ and $G_n^{(2,\frac{1}{2})}$ outperform F_n
- $G_n^{(1,\frac{1}{2})}$ with $a = -3$ and $b = 3$ always slightly better than $G_n^{(2,\frac{1}{2})}$
- $G_n^{(j,0)}$ and $G_n^{(j,1)}$, $j = 1, 2$ have irregular behaviour: better or worse than F_n depending on n and the simulated distribution. Nevertheless, their estimated MISE are always greater than $G_n^{(j,\frac{1}{2})}$
- $G_n^{(1,\frac{1}{2})}$ outperforms both K_n and F_n for 6/16 tested distributions (from only $n = 50$ for 2 of them)

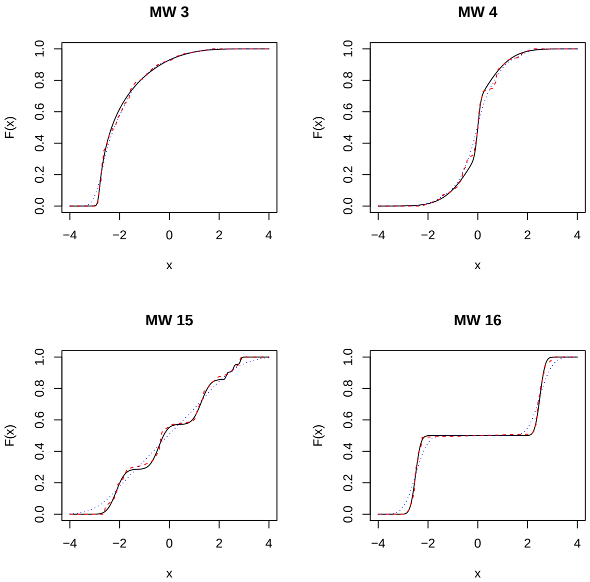
Results

- For all tested distributions and all n in $\{20, 50, 100\}$, $G_n^{(1,\frac{1}{2})}$ and $G_n^{(2,\frac{1}{2})}$ outperform F_n
- $G_n^{(1,\frac{1}{2})}$ with $a = -3$ and $b = 3$ always slightly better than $G_n^{(2,\frac{1}{2})}$
- $G_n^{(j,0)}$ and $G_n^{(j,1)}$, $j = 1, 2$ have irregular behaviour: better or worse than F_n depending on n and the simulated distribution. Nevertheless, their estimated MISE are always greater than $G_n^{(j,\frac{1}{2})}$
- $G_n^{(1,\frac{1}{2})}$ outperforms both K_n and F_n for 6/16 tested distributions (from only $n = 50$ for 2 of them)

Number	Name	Distribution function: $\sum_{\ell=0}^{k} \omega_\ell \mathcal{N}(\mu_\ell, \sigma_\ell^2)$
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	$\frac{1}{10}\mathcal{N}(0,1) + \frac{9}{10}\mathcal{N}(0,(\frac{1}{10})^2)$
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: Selected distribution functions used in the simulation study: #1-#15 are from MW 92, #16 from JMVS95


MW 3 (strongly skewed)
MW 4 (kurtotic unimodal)
MW 15 (discrete comb)
MW 16 (distant bimodal)
f(x)
x
0.0
0.4
0.8
1.2
0.5
1.0
1.5
0.1
0.2
0.3
−3
−2
−1
0
1
2
3


MW 3
MW 4
MW 15
MW 16
F(x)
x
0.0 0.2 0.4 0.6 0.8 1.0
−4 −2 0 2 4

- N. Altman and C. Léger.Bandwidth selection for kernel distribution estimation *J. Statist. Plann. Infer.*, 46(2):195–214, 1995.
- 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.
- H. A. David and H. N. Nagaraja. *Order statistics*. Wiley Series in Probability and Statistics. Third edition, 2003.
- J. S. Marron and M. P. Wand. Exact mean integrated squared error. *Ann. Statist.*, 20(2):712–736, 1992.
- 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.
- R. R. Read. The asymptotic inadmissibility of the sample distribution function. *Ann. Math. Statist.*, 43:89–95, 1972.
- P. Sarda. Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference*, 35(1):65–75, 1993.
- R. Servien. Estimation de la fonction de répartition : revue bibliographique. *J. SFdS*, 150(2):84–104, 2009.