Distribution estimator defined by Lagrange polynomials

Salima Helali and Yousri Slaoui
Sfax University and Poitiers University
helali.salima@gmail.com
Yousri.Slaoui@math.univ-poitiers.fr

1 Introduction

Let $X_1, \dots, X_n$ be a sequence of independent identically distributed (i.i.d.) random variables having a common unknown distribution function F with associated density f supported on $[-1, 1]$.
We considered in this paper an estimator of order $m > 0$ of the distribution F using **Lagrange polynomial** and **Tchebytchev's points**, defined as

$$\tilde{F}_{n,m}(x) := \sum_{i=1}^{m} \widehat{F}_n(x_i)\mathcal{L}_i(x), \tag{1}$$

where, for all $i = 1 \dots m$, $x_i = \cos\left(\frac{(2i-1)\pi}{2m}\right)$ are Tchebytchev's discretization points, $\mathcal{L}_i(x) = \prod_{j=1, j\neq i}^{m} \frac{x - x_j}{x_i - x_j}$ is the Lagrange polynomial, and $\widehat{F}_n(x) = \sum_{l=1}^{n} \mathbb{1}_{\{X_l \leq x\}}$ denotes the empirical distribution function obtained from a random sample of size n and assuming that $m = m_n$ (depends on n).
The aim of this work is to study the properties of the distribution estimator (1), as a competitor for **Vitale**'s distribution estimator **(1975)** defined by

$$\overline{F}_{n,\nu}(x) := \sum_{k=0}^{\nu} \widehat{F}_n\left(\frac{k}{\nu}\right) b_k(\nu, x), \tag{2}$$

with $\widehat{F}_n$ is the empirical distribution function and $b_k(\nu, x) = C_\nu^k x^k (1-x)^{\nu-k}$ is the Bernstein polynomial of order $\nu > 0$ and assuming that $\nu = \nu_n$ (depends on n).

2 Preliminaries and Notations

We define the following class of regularly varying sequences.

Let $\gamma \in \mathbb{R}$ and $(v_n)_{n\geq 1}$ be a nonrandom positive sequence. We say that $(v_n) \in \mathcal{GS}(\gamma)$ if

$$\lim_{n\to+\infty} n\left[1 - \frac{v_{n-1}}{v_n}\right] = \gamma.$$

This condition was introduced by **Galambos** and **Seneta (1973)** to defne regularly varying sequences.

Throughout this paper, we will use the following notations: Let $m \geq 1$, $i = 1 \dots m$ and $x \in [-1, 1]$, we note $\theta_i = \frac{(2i-1)\pi}{2m}$ and $x_i = \cos(\theta_i)$: Tchebytchev's discretization points,
$\mathcal{L}_i(x) = \prod_{j=1, j\neq i}^{m} \frac{x - x_j}{x_i - x_j}$: Lagrange polynomial, $A_m(x) = \sum_{i=1}^{m} F(x_i)\mathcal{L}_i(x)$, $\| g \| = \sup_{-1\leq x\leq 1} | g(x) |$,
$T_m(x) = \prod_{i=1}^{m}(x - x_i)$: Tchebytchev polynomial, $J_m(x) = \sum_{k=1}^{m} | x_k - x | \mathcal{L}_k^2(x)$, $\sigma^2(x) = F(x)(1 - F(x)$,
$S_m(x) = \sum_{k=1}^{m} \mathcal{L}_k^2(x)$, $P_m(x) = \sum\sum_{0\leq k<l\leq m} (x_k - x)\mathcal{L}_k(x)\mathcal{L}_l(x)$.

3 Results

Hypotheses:

To obtain the behaviour of the estimator defined in equation (1) inside the interval $[-1, 1]$, we make the assumption that

$(\mathcal{A}_1)$ F is continuous and admits two continuous and bounded derivatives on $[-1, 1]$.
$(\mathcal{A}_2)$ $(\nu_n) \in \mathcal{GS}(a)$, $(m_n) \in \mathcal{GS}(a)$, $a \in (0, 1)$.

Proposition 1

Under assumption $(\mathcal{A}_1)$, we have for $x \in [-1, 1]$ that

$$Bias(\tilde{F}_{n,m}) = \mathbb{E}(\tilde{F}_{n,m}) - F(x) = \frac{\pi}{2}T_m(x)m^{-2}f(x) + o(m^{-4}), \tag{3}$$
$$Var(\tilde{F}_{n,m}) = n^{-1}\sigma^2(x) + 2f(x)P_m(x)n^{-1} + n^{-1}O(J_m(x)) + O(n^{-1}m^{-4}). \tag{4}$$

Notice that the previous result implies that the bias of $\tilde{F}_{n,m}$ is $O(m^{-4})$ which is smaller than the bias of estimator obtained using Bernstein polynomial having a bias as $O(m^{-1})$. The following proposition shows that $\tilde{F}_{n,m}$ is strongly consistent.

Proposition 2

Under assumption $(\mathcal{A}_1)$, if $n, m \to \infty$, then $\| \tilde{F}_{n,m} - F \| \to 0$ almost surely (a.s.).

Finally, the following proposition shows the asymptotic normality of the estimator (1).

Proposition 3

Assume $(\mathcal{A}_1)$ holds and $m, n \to \infty$. For $x \in (-1, 1)$, we have that

$$n^{1/2}\left(\tilde{F}_{n,m}(x) - A_m(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(x)). \tag{5}$$

4 Numerical studies

In this section, under assumption $(\mathcal{A}_2)$, we consider **the cross-validation** procedure, which is a usual way of selecting the smoothing parameter (m_n) of the proposed estimator (1) and (ν_n) of Vitale's estimator (2). **Sarda (1993)** proposed to use

$$CV(m) = \sum_{i=1}^{n}\left(\widehat{F}_n(x_i) - F_{-i}(x_i)\right)^2.$$

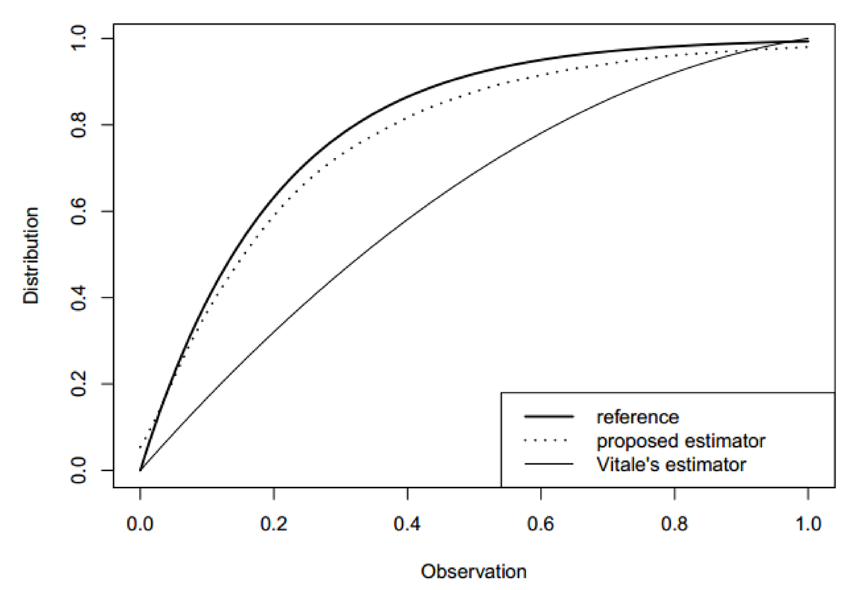

Figure 1: Qualitative comparison between Vitale's estimator defined in (2) and the proposed distribution estimator (1), for 500 samples of size 50 for the beta distribution $\mathcal{E}(5)$.

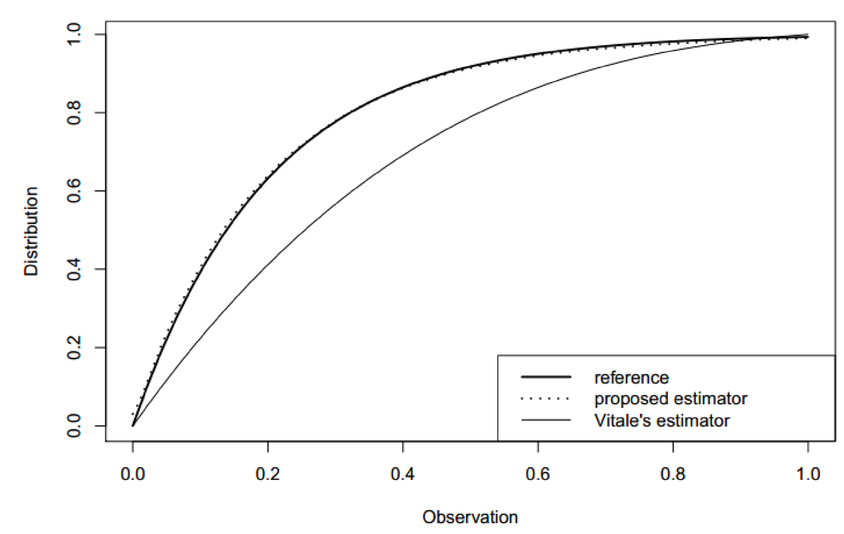

Figure 2: Qualitative comparison between Vitale's estimator defined in (2) and the proposed distribution estimator (1), for 500 samples of size 100 for the beta distribution $\mathcal{E}(5)$.

Lois	n (size)	Proposed estimator	Vitale's estimator
$\mathcal{B}(10, 10)$	10	**0.03694811**	0.03299928
	30	**0.01894400**	0.03299865
	50	**0.01401786**	0.03266988
	100	**0.00993235**	0.02615096
$\mathcal{B}(2, 2)$	10	0.01057271	**0.00474892**
	30	**0.00271338**	0.00471605
	50	**0.00127222**	0.00466732
	100	**0.00059276**	0.00341529
$\mathcal{E}(5)$	10	**0.02272625**	0.06357428
	30	**0.00403700**	0.04706539
	50	**0.00109249**	0.04232919
	100	**4.95561e^{-5}**	0.01800240

Table 1: ISE (Integrated Squared Error) for $N = 500$ samples of Vitale's estimator and the proposed estimator $\tilde{F}_{n,m}$

5 Conclusion

In this work, we propose an estimator of the distribution function using Lagrange polynomials and Tchebytchev'points. We study its asymptotic behaviours. Then, we compare our proposed estimator to the Vitale's distribution estimator throught simulations and computing the ISE. In conclusion, using the proposed estimator $\tilde{F}_{n,m}$ we can obtain better results than those given by Vitale's distribution estimator. Hence, we plan to work on improvement of this estimator using stochastic algorithm and then we can compare the proposed work to the one given in Slaoui (2014b) and Jmaei et al. (2017).

References

[1] Jmaei, A., Slaoui, Y., and Dellagi, W. (2017). Recursive distribution estimator defined by stochastic approximation method using Bernstein polynomials. Journal of Nonparametric Statistics, 29(4), 792-805.

[2] Serfling, R. J. (2009). Approximation theorems of mathematical statistics (Vol. 162). John Wiley and Sons.

[3] Slaoui, Y. (2014b). The stochastic approximation method for estimation of a distribution function. Mathematical Methods of Statistics, 23(4), 306-325.

[4] Vitale, R. A. (1975). A Bernstein polynomial approach to density function estimation. In Statistical inference and related topics (pp. 87-99).