# Estimation of a distribution function defined by Lagrange polynomials and Tchebytchev points

Salima HELALI [1] & Yousri SLAOUI [2]

[1] *helali.salima@gmail.com*
[2] *yousri.slaoui@univ-poitiers.fr*

**Résumé.** Nous proposons un estimateur d'une fonction de répartition en utilisant le polynôme d'interpolation de Lagrange et les points de Tchebytchev. Nous étudions les propriétés de cet estimateur et nous les comparons avec celle de l'estimateur d'une fonction de répartition de Vitale. Nous montrons que notre estimateur domine celui de Vitale en terme de risque. Ensuite, nous confirmons ces résultats théoriques par des simulations.

**Mots-clés.** Estimation d'une fonction de répartition, Polynôme de Lagrange, Polynôme de Tchebytchev, Polynôme de Bernstein.

**Abstract.** We consider an application of Lagrange polynomials and Tchebytchev's points for estimating a distribution function with support $[-1, 1]$. We study the properties of this estimator, as a competitor of Vitale's distribution estimator defined by Bernstein polynomials. We show that, this estimator dominates Vitale's estimator in terms of risk. Finally, we confirm our theoretical results through a simulation study.

**Keywords.** Distribution estimator, Lagrange polynomials, Tchebytchev's points, Bernstein polynomials, Asymptotic properties.

# 1 Introduction

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables having a common unknown distribution function $F$ with associated density $f$ supported on $[-1, 1]$. Now, since we know that $F$ is continuous, we consider the estimation of $F$ by using smooth functions rather than the empirical distribution function, which is not continuous. There have been several methods for smooth estimation of density and distribution functions such as kernel methods. However, these methods have estimation problems at the edges, when we have a random variable $X$ with distribution $F$ supported on a compact interval. In order to solve this problem, there have been many methods such as the Bernstein polynomial distribution estimator proposed first by Vitale (1975) and then extended by Babu et al. (2002). In this short communication, we suppose that $f$ is supported on $[-1, 1]$, and we propose the following estimator of order $m > 0$ of the distribution $F$ using Lagrange polynomial, 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 $i = 1 \dots m$, $x_i = \cos\left(\dfrac{(2i-1)\pi}{2m}\right)$ are Tchebytchev's points,

$\mathcal{L}_i(x) = \displaystyle\prod_{j=1, j\neq i}^{m} \frac{x - x_j}{x_i - x_j}$ is the Lagrange polynomial and $\widehat{F}_n$ denotes the empirical distribution function obtained from a random sample of size $n$. We assume that $m = m_n$ (depends on $n$). The aim of this short communication 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 $\nu > 0$. We assume that $\nu = \nu_n$ (depends on $n$).

# 2 Assumptions and notations

We define the following class of regularly varying sequences.

**Definition 1.** *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 define regularly varying sequences.

To study the asymptotic behaviours of the estimator (1) inside the interval $[-1, 1]$, we make the following assumptions:

$(\mathcal{A}_1)$ $F$ is of class $C^2$ on $[-1, 1]$.

$(\mathcal{A}_2)$ $(\nu_n) \in \mathcal{GS}(a)$, $(m_n) \in \mathcal{GS}(a)$, $a \in (0, 1)$.

The assumption $(\mathcal{A}_1)$ is used in the theoretical part to calculate the bias and the variance of $\tilde{F}_{n,m}$. The assumption $(\mathcal{A}_2)$ is used in the numerical studies to obtain the optimal choices of the orders $(m_n)$ and $(\nu_n)$ which minimize the risk. Throughout this communication, we will use the following notations for $m \geq 1$, $i = 1 \dots m$ and $x \in [-1, 1]$:

$\theta_i = \dfrac{(2i-1)\pi}{2m}$ and $x_i = \cos(\theta_i)$: Tchebytchev's points,

$T_m(x) = \displaystyle\prod_{i=1}^{m}(x - x_i)$: Tchebytchev polynomial, $J_m(x) = \sum_{k=1}^{m} \mid x_k - x \mid \mathcal{L}_k^2(x)$,

$\mathcal{L}_i(x) = \displaystyle\prod_{j=1, j\neq i}^{m} \frac{x - x_j}{x_i - x_j}$: Lagrange polynomial, $\sigma^2(x) = F(x)(1 - F(x))$,

$S_m(x) = \displaystyle\sum_{k=1}^{m} \mathcal{L}_k^2(x)$, $P_m(x) = \sum_{0\leq k<l\leq m}\sum (x_k - x)\mathcal{L}_k(x)\mathcal{L}_l(x)$, $A_m(x) = \displaystyle\sum_{i=1}^{m} F(x_i)\mathcal{L}_i(x)$.

# 3 Main results

Our first result is the following proposition which gives the bias and the variance of $\tilde{F}_{n,m}$.

**Proposition 1** (**Bias and variance of** $\tilde{F}_{n,m}$)**.** *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 the 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** (**Uniform convergence of** $\tilde{F}_{n,m}$)**.** *Under assumption* $(\mathcal{A}_1)$*, if* $n, m \to \infty$*, then* $\left\|\tilde{F}_{n,m} - F\right\| \to 0$ *almost surely (a.s.), where* $\|G\| = \sup_{x\in[-1,1]} |G(x)|$ *for any bounded function* $G$ *on* $[-1,1]$.

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

**Proposition 3** (**Asymptotic normality of** $\tilde{F}_{n,m}$ )**.** *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, we show that the estimator using the Lagrange polynomial defined in equation (1) outperformed the estimator using the Bernstein polynomial defined in equation (2). In our simulation study, we consider three sample sizes, $n = 30$, $n = 50$, $n = 100$ and the following two distribution functions:

*a*) the beta distribution $\mathcal{B}(2,2)$.

*b*) the exponential distribution $\mathcal{E}(5)$.

In the framework of the nonparametric estimators, the smoothing parameter selection methods studied in the literature can be divided into three broad classes: the cross-validation techniques, the plug-in methods and the bootstrap idea. In this section, we assume that the assumption $(\mathcal{A}_2)$is satisfied. We use the cross-validation procedure to select respectively the smoothing order $(m_n)$ of the proposed estimator (1) and the order $(\nu_n)$ of the Vitale's estimator (2). Sarda (1993) proposed to use

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

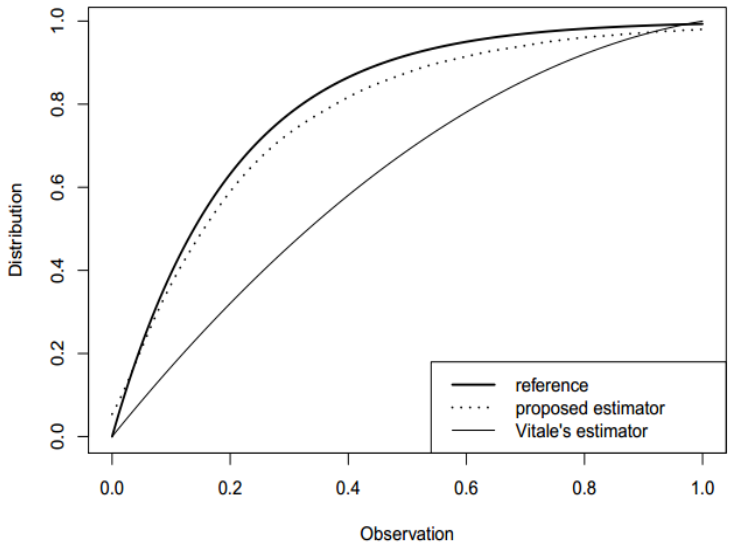


Figure 1: Qualitative comparison between the estimator $\tilde{F}_{n,m}$ defined in (2) and the proposed distribution estimator (1), for 500 samples of size 50 for the exponential distribution $\mathcal{E}(5)$.

For each distribution function and sample size $n$, we compute the Integrated Squared Error ($ISE$) of the estimator over $N = 500$ trials,

$$ISE[\widehat{g}] = \int_0^1 (\widehat{g}(x) - F(x))^2 \, dx,$$

where $\widehat{g}$ is an estimator of the distribution $F$.

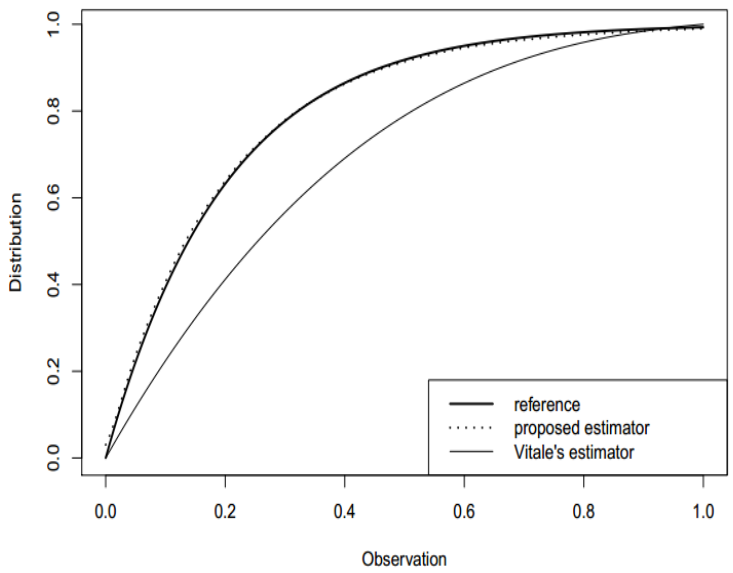


Figure 2: Qualitative comparison between the estimator $\tilde{F}_{n,m}$ defined in (2) and the proposed distribution estimator (1), for 500 samples of size 100 for the exponential distribution $\mathcal{E}(5)$.

Table 1: $ISE$ for $N = 500$ trials of the Vitale's estimator and the proposed estimator $\tilde{F}_{n,m}$

| **laws** | $n$ | **Proposed estimator** | **Vitale's estimator** |
|---|---|---|---|
| $(a)$ | 30 | **0.00271338** | 0.00471605 |
| | 50 | **0.00127222** | 0.00466732 |
| | 100 | **0.00059276** | 0.00341529 |
| $(b)$ | 30 | **0.00403700** | 0.04706539 |
| | 50 | **0.00109249** | 0.04232919 |
| | 100 | $\mathbf{4.95561e^{-5}}$ | 0.01800240 |

From figures 1, 2 and table 1, we conclude that

- In all the considered distributions, by choosing the appropriate $(m_n)$, the $ISE$ of the distribution estimator (1) is smaller than that of Vitale's estimator (2).
- The $ISE$ decreases as the sample size increases.

# 5 Conclusion

In this communication, we propose an estimator of a distribution function using Lagrange polynomials and Tchebytchev's points. We study its asymptotic behaviours. The proposed estimator is asymptotically normal. Then, we compare our proposed estimator to the Vitale's distribution estimator through a simulation study. For all the considered cases, the $ISE$ of our proposed estimator (1) is smaller than that of Vitale's estimator (2). 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 make an extension of the current work by considering a recursive version and to compare the obtained estimators to the one given in Slaoui (2014b) and Jmaei et al. (2017). We plan also to consider the estimation of a density function in a recursive framework (see Slaoui (2014a)) and then the estimation of a regression function in a recursive framework by using Lagrange polynomials (see Slaoui (2015, 2016)).

# Bibliographie


Babu, G.J., Canty, A.J., and Chaubey, Y.P. (2002), Application of Bernstein Polynomials for Smooth Estimation of a Distribution and Density Function, *Journal of Statistical Planning and Inference*, 105, 377–392.

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), 792805.

Leblanc, A. (2012). On estimating distribution functions using Bernstein polynomials, *Annals of the Institute of Statistical Mathematics*, 64(5), 919943.

Mokkadem, A., Pelletier, M., and Slaoui, Y. (2009), The Stochastic Approximation Method for the Estimation of a Multivariate Probability Density, *Journal of Statistical Planning and Inference*, 139, 24592478.

Sarda, P. (1993), Smoothing Parameter Selection for Smooth Distribution Functions, *Journal of Statistical Planning and Inference*, 35, 65–75.

Slaoui, Y. (2014a). Bandwidth selection for recursive kernel density estimators defined by stochastic approximation method, *Journal of Probability and Statistics*, 2014, ID 739640, doi:10.1155/2014/739640.

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

Slaoui, Y. (2015). Plug-In Bandwidth selector for recursive kernel regression estimators defined by stochastic approximation method, *Statistica Neerlandica*, 69(4), 483–509.

Slaoui, Y. (2016). Optimal bandwidth selection for semi-recursive kernel regression estimators, *Statistics and Its Interface*, 9(3), 375388.

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