Semi-Parametric Estimation Using Bernstein Polynomial and a Finite Gaussian Mixture Model

Salima Helali, Afif Masmoudi, Yousri Slaoui

entropy

MDPI

Article

Semi-Parametric Estimation Using Bernstein Polynomial and a Finite Gaussian Mixture Model

Salima Helali [1], Afif Masmoudi [2] and Yousri Slaoui [3,*]

[1] Mathematics Laboratory, Angers University, 49100 Angers, France; salima.helali@univ-angers.fr
[2] Probability and Statistics Laboratory, Sfax University, 3029 Sfax, Tunisia; afif.masmoudi@fss.usf.tn
[3] Mathematics and Applications Laboratory, Poitiers University, 86073 Poitiers, France
* Correspondence: yousri.slaoui@math.univ-poitiers.fr

Abstract: The central focus of this paper is upon the alleviation of the boundary problem when the probability density function has a bounded support. Mixtures of beta densities have led to different methods of density estimation for data assumed to have compact support. Among these methods, we mention Bernstein polynomials which leads to an improvement of edge properties for the density function estimator. In this paper, we set forward a shrinkage method using the Bernstein polynomial and a finite Gaussian mixture model to construct a semi-parametric density estimator, which improves the approximation at the edges. Some asymptotic properties of the proposed approach are investigated, such as its probability convergence and its asymptotic normality. In order to evaluate the performance of the proposed estimator, a simulation study and some real data sets were carried out.

Keywords: asymptotic properties; Bernstein polynomial; EM algorithm; Gaussian mixture model; kernel estimator; shrinkage estimator

Citation: Helali, S.; Masmoudi, A.; Slaoui, Y. Semi-Parametric Estimation Using Bernstein Polynomial and a Finite Gaussian Mixture Model. *Entropy* **2022**, *24*, 315. https://doi.org/10.3390/e24030315

Academic Editor: Udo Von Toussaint

Received: 17 January 2022
Accepted: 17 February 2022
Published: 23 February 2022

1. Introduction

Density estimation is a widely adopted tool for multiple tasks in statistical inference, machine learning, visualization and exploratory data analysis. Existing density estimation algorithms can be categorized into either parametric, semi-parametric, or non-parametric approaches. In the non-parametric framework, several methods have been set forward for the smooth estimation of density and distribution functions. The most popular one, called kernel method, was introduced by [1]. The advances were carried out by [2] to estimate a density function. The reader is recommended to consult the paper [3] for an introduction to several kernel smoothing techniques. However, kernel methods display estimation problems at the edges, when we have a random variable X with density function f supported on a compact interval. Moreover, if $X_1, \ldots, X_n$ is a sample with the same density f, it is well known, in non-parametric kernel density estimation, that the bias of the standard kernel density estimator

$$\widehat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{1}$$

is of a larger order near the boundary than that in the interior, where K is a kernel (that is, a positive function satisfying $\int K(x)dx = 1$) and (h_n) is a bandwidth (that is, a sequence of positive real numbers that goes to zero). Let us now suppose that f has two continuous derivatives everywhere and that, as $n \to \infty$, $h = h_n \to 0$ and $nh \to 0$. Let $x = ph$ for $p > 0$. Near the boundary, the expression of the mean and the variance are indicated as

$$\mathbb{E}\left[\widehat{f}_n(x)\right] \simeq f(x)\int_{-\infty}^{p} K(x)dx - f'(x)\int_{-\infty}^{p} xK(x)dx + \frac{1}{2}h^2 f''(x)\int_{-\infty}^{p} x^2K(x)dx,$$

and

$$Var\left[\widehat{f}_n(x)\right] \simeq (nh)^{-1} f(x) \int_{-\infty}^{p} K^2(x)dx.$$

These bias phenomena are called boundary bias. Numerous authors have elaborated methods for reducing these phenomena, such as data reflection [4], boundary kernels [5–7], local linear estimator [8,9], use of beta and gamma kernels [10,11] and bias reduction [12,13]. For a smooth estimator of a density function f with finite known support, there have been several methods, such as Vitale's method [14], which is based on Bernstein polynomials and expressed as

$$\tilde{f}_{1,n,m}(x) = m \sum_{k=0}^{m-1} \left[F_n\left(\frac{k+1}{m}\right) - F_n\left(\frac{k}{m}\right) \right] b_k(m-1,x), \quad (2)$$

where F_n is the empirical distribution function and $b_k(m,x) = C_m^k x^k (1-x)^{m-k}$ is the Bernstein polynomial. This estimator was investigated in the literatures [15–18] and, more recently, by [12,19,20].

Within the parametric framework, it is noteworthy that the Gaussian mixture model can be used to estimate any density function, without any problem of estimation on the edge. This refers to the fact that the set of all normal mixture densities is dense in the set of all density functions under the L^1 metric [21]. The investigation of mixture models stands for a full field in modern statistics. It is a probabilistic model introduced by [22] to illustrate the presence of subpopulations within an overall population. It has been developed so far by various authors, such as [23]. It is used for data classification and it provides efficient approaches of model-based clustering. The authors of [24] demonstrated that, when a Gaussian mixture model is used to estimate a density non-parametrically, the density estimator that uses the Bayesian information criterion (BIC) of [25] to select the number of components in the mixture is consistent [26].

However, we obtain the non-parametric kernel estimate of a density if we fit a mixture of n components in equal proportions $1/n$, where n is the size of the observed sample. As a matter of fact, it can be inferred that mixture models occupy an interesting niche between parametric and non-parametric approaches to statistical estimation.

More recently, in the parametric context, [27] proposed a parametric model using Bernstein polynomials with positive coefficients to estimate the unknown density function f; this estimator is defined as follows:

$$f_B(x, p_m) = \sum_{i=1}^{m} \widehat{p}_{mi} \mathcal{B}_{mi}(x), \quad (3)$$

where $B_{mi}(x) = (m+1)b_i(m,x)$, for $i = 0, \ldots, m$, $p_m = (p_{m1}, \ldots, p_{mm})^T$ ($p_{mi} \geq 0$, $i = 1, \ldots, m$, $\sum_{i=1}^{m} p_{mi} \leq 1$) and $\widehat{p}_{mi}$ are the estimators of the parameters p_{mi}, obtained by the Expectation Maximization (EM) algorithm as follows:

$$p_{mi}^{(s+1)} = \frac{1}{n} \sum_{j=1}^{n} \frac{p_{mi}^{(s)} \mathcal{B}_{mi}(x_j)}{\sum_{k=0}^{m} p_{mk}^{(s)} \mathcal{B}_{mk}(x_j)}, \; i = 0, \ldots, m; \; s = 0, 1, \ldots$$

with $\widehat{p}_{mi} = \lim_{s \to \infty} p_{mi}^{(s)}$, for $i = 1, \ldots, m$. The proposed method gives a consistent estimator in L^2 distance under some conditions.

The problem at the edge does not arise for the parametric model. For this reason, the basic idea of this work is to consider a shrinkage method using Bernstein (Vitale's estimator) and Gaussian mixture estimators, to construct a shrinkage density estimator, in order to improve the approximation at the edge. A shrinkage estimator is a convex combination between estimators [28]. Basically, this implies that a naive or raw estimate is improved by combining it with other information.

The remainder of this paper is organized as follows: In the next section, we recall some intrinsic properties of the classical EM algorithm in the context of the Gaussian mixture parameter estimation. In Section 3, we introduce a new semi-parametric estimation approach based on the shrinkage method using Bernstein polynomials and Gaussian mixture densities. In Section 4, the consistency of the proposed estimator is exhibited, as well as its asymptotic normality.

Section 5 highlights a simulation study that compares the performance of the proposed approach with the Bernstein estimator, the standard Gaussian kernel estimator and Guan's estimator. The closing Section 6 crowns the whole work, wraps the conclusion and provides new perspectives for future work.

2. Background

The Gaussian Mixture Model and Em Algorithm

Let us consider $X = (X_1, \ldots, X_n)$, a sequence of independent and identically distributed (i.i.d.) with common Gaussian mixture density defined by

$$g(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k)(x), \tag{4}$$

where

$$\theta = (\pi, \mu, \sigma) = (\pi_1 \ldots, \pi_K, \mu_1 \ldots, \mu_K, \sigma_1, \ldots, \sigma_K),$$

satisfies

$$0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_k = 1, \mu_k \in \mathbb{R}, \sigma_k > 0, \text{ for } k = 1, \ldots, K \text{ where } K > 0,$$

and

$$\mathcal{N}(\mu, \sigma)(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Finally, for each observed data point X_i, we associate a component label vector Z_i in order to manage the data clustering. This random vector $Z_i = (Z_{ik})_{1 \leq k \leq K}$ is defined such that $Z_{ik} = 1$ if the considered observation X_i is drawn form the k^{kt} component of the mixture and $Z_{ik} = 0$ otherwise. Consequently, Z_i is distributed as a multivariate Bernoulli distribution with vector parameters $(\pi_1, \ldots, \pi_K)$ as follows:

$$\mathbb{P}(Z_i = z_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}}.$$

The EM algorithm is a popular tool in statistical estimation problems involving *incomplete data* or problems which can be posed in a similar form, such as the mixture parameters estimation [23,29]. In the EM framework, $(X_1, \ldots, X_n, Z_1, \ldots, Z_n)$ corresponds to the complete data and $(Z_1, \ldots, Z_n)$ stand for the hidden data. Hence, the complete-data log-likelyhood is expressed by

$$L(X_1, \ldots, X_n, Z_1, \ldots, Z_n, \theta) = \sum_{i=1}^{n} \sum_{j=1}^{K} Z_{ij} \left[\log(\pi_j) + \log\left(\mathcal{N}(\mu_j, \sigma_j)(X_i)\right)\right]. \tag{5}$$

The two steps of the EM algorithm, after l iterations, are the following:

(i) *E-step*: The conditional expectation of the complete-data log-likelyhood given the observed data, using the current fit $\theta^{(l)}$, is defined by

$$\varphi\left(\theta|\theta^{(l)}\right) = \mathbb{E}_{\theta^{(l)}}(L(X_1, \ldots, X_n, Z_1, \ldots, Z_n, \theta)|X_1, \ldots, X_n). \tag{6}$$

The posterior probability that X_i belongs to the jth component of the mixture at the lth iteration, is expressed as

$$\tau_{ij}^{(l)} = \mathbb{E}_{\theta^{(l)}}\left(Z_{ij}|X_1,\ldots,X_n\right) = \frac{\pi_j^{(l)}\mathcal{N}\left(\mu_j^{(l)},(\sigma^2)_j^{(l)}\right)(X_i)}{\sum_{h=1}^{K}\pi_h^{(l)}\mathcal{N}\left(\mu_h^{(l)},(\sigma_h^2)^{(l)}\right)(X_i)}. \tag{7}$$

Finally, we obtain

$$\varphi\left(\theta|\theta^{(l)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{K}\tau_{ij}^{(l)}\left[\log(\pi_j)+\log(\mathcal{N}(\mu_j,\sigma_j)(X_i))\right]. \tag{8}$$

(ii) *M-step*: It consists of a global maximization of $\varphi\left(\theta|\theta^{(l)}\right)$ with respect to θ.

$$\theta^{(l+1)} = \arg\max_{\theta}\varphi\left(\theta|\theta^{(l)}\right). \tag{9}$$

The updated estimates are stated by

$$\pi_j^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ij}^{(l)}, \tag{10}$$

$$\mu_j^{(l+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(l)}X_i}{\sum_{i=1}^{n}\tau_{ij}^{(l)}}, \tag{11}$$

$$(\sigma_j^2)^{(l+1)} = \frac{\sum_{i=1}^{n}\tau_{ij}^{(l)}\left(X_i-\mu_j^{(l+1)}\right)^2}{\sum_{i=1}^{n}\tau_{ij}^{(l)}}. \tag{12}$$

We repeat these two steps until $\left|\left|\theta^{(l+1)}-\theta^{(l)}\right|\right| < \epsilon$, where ϵ is a fixed threshold of convergence. The convergence properties of the EM algorithm have been investigated by [29] and by [30]. Relying upon Jensen's inequality, it can be noticed that, as $\varphi\left(\theta|\theta^{(l)}\right)$ increases, the log-likelihood function also increases [29]. Consequently, the EM algorithm converges within a finite iteration number and gives the parameters' maximum likelihood estimates. Therefore, under some conditions and according to [29], we have

$$\lim_{l\to\infty}\pi_j^{(l)} = \widehat{\pi}_j,\ \lim_{l\to\infty}\mu_j^{(l)} = \widehat{\mu}_j \text{ and } \lim_{l\to\infty}(\sigma_j^2)^{(l)} = \widehat{\sigma^2}_j \text{ almost surely (a-s).} \tag{13}$$

In what follows, $\widehat{\theta} = (\widehat{\pi}_1,\ldots,\widehat{\pi}_K,\widehat{\mu}_1,\ldots,\widehat{\mu}_K,\widehat{\sigma}_1,\ldots,\widehat{\sigma}_K)$.

3. Proposed Approach

The proposed semi-parametric approach rests upon the *shrinkage* combination between the Gaussian mixture model and the Bernstein density estimators using the EM algorithm for the parameter estimations. The literature on shrinkage estimation is enormous. From this perspective, it is noteworthy to mention the most relevant contributions. The authors of [28] were the first to introduce the classic shrinkage estimator. The authors of [31] provided theory for the analysis of risk. Oman [32,33] developed estimators which shrink Gaussian density estimators towards linear subspaces. An in-depth investigation of shrinkage theory is displayed in Chapter 5 of [34].

The proposed semi-parametric approach based upon estimating the density function f relies on the same principle of Stein's works and there are two aspects along this line. The first setting is non-parametric in the sense that we do not assume any parametric form

of the density. The non-parametric setting is very important as it allows us to perform statistical inference without making any assumption on the parametric form of the true density f. The second setting is to consider the Gaussian mixture model as a parametric estimator of the unknown density f.

In what follows, we consider $X_1, \ldots, X_n$ a sequence of i.i.d. random variables having a common unknown density function f supported on $[0,1]$. We here develop a shrinkage method to estimate the density function, which is divided into the following three steps:

Step 1 We consider the Bernstein estimator of the density function f, which is defined as

$$\tilde{f}_{1,n,m}(x) = m \sum_{i=0}^{m-1} \left[F_n\left(\frac{i+1}{m}\right) - F_n\left(\frac{i}{m}\right) \right] b_i(m-1,x) \tag{14}$$

Step 2 In view of (13), we consider the Gaussian mixture density as an estimator of the density function f, given by

$$\tilde{f}_{2,n}(x) = \sum_{k=1}^{K} \widehat{\pi}_k \mathcal{N}(\widehat{\mu}_k, \widehat{\sigma}_k)(x), \tag{15}$$

where $\widehat{\mu}_k$, $\widehat{\sigma}_k$ and $\widehat{\pi}_k$ are estimated by the EM algorithm defined in (13).

Step 3 We consider the shrinkage density estimator $\widehat{f}_{n,m}$ form defined by

$$\widehat{f}_{n,m}(x) = \lambda \tilde{f}_{1,n,m}(x) + (1-\lambda)\tilde{f}_{2,n}(x),$$

and we use the EM algorithm to estimate the parameter $\lambda \in [0,1]$ of the proposed model.

By the same way as considered in Section 2, the two steps of the EM algorithm, after t iterations, are denoted in terms of the following:

1. *E-step*: The conditional expectation of the complete-data log-likelihood given the observed data, using the current $\lambda^{(t)}$, is provided by

$$Q(\lambda|\lambda^{(t)}) = \sum_{i=1}^{n} \mathbb{E}_{\lambda^{(t)}}(W_{i1} \mid X_i) \log \tilde{f}_{1,n,m}(X_i) + \mathbb{E}_{\lambda^{(t)}}(W_{i2} \mid X_i) \log \tilde{f}_{2,n}(X_i),$$

where $W_i = (W_{i1}, W_{i2})$ is a discrete random vector, following a multivariate Bernoulli distribution with vector parameters $(\lambda, 1-\lambda)$. Using Bayes's formula, we obtain the posterior probability in the tth iteration denoted by

$$\overline{\tau}_{i1}^{(t)} = \frac{\tilde{f}_{1,n,m}(X_i)\lambda^{(t)}}{\lambda^{(t)}\tilde{f}_{1,n,m}(X_i) + (1-\lambda^{(t)})\tilde{f}_{2,n}(X_i)},$$

and

$$\overline{\tau}_{i2}^{(t)} = \frac{\tilde{f}_{2,n}(X_i)\lambda^{(t)}}{\lambda^{(t)}\tilde{f}_{1,n,m}(X_i) + (1-\lambda^{(t)})\tilde{f}_{2,n}(X_i)} = 1 - \overline{\tau}_{i1}^{(t)}.$$

2. *M-step*: It consists of a global maximization of $Q(\lambda|\lambda^{(t)})$ with respect to λ.

$$\lambda^{(t+1)} = \arg\max_{\lambda} Q(\lambda \mid \lambda^{(t)}).$$

The updated estimate of λ is indicated by

$$\lambda^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \overline{\tau}_{i1}^{(t)}.$$

The estimation of λ is obtained from by iterating the EM algorithm until convergence.

$$\lim_{t\to\infty} \lambda^{(t)} = \widehat{\lambda}. \tag{16}$$

Therefore, the proposed estimator of the density function f is defined by

$$\widehat{f}_{n,m}(x) \quad = \quad \widehat{\lambda}\tilde{f}_{1,n,m}(x) + (1-\widehat{\lambda})\tilde{f}_{2,n}(x). \tag{17}$$

Basically, it is a shrinkage estimator that shrinks the Bernstein estimator towards the Gaussian mixture density by a specified amount of λ. If $\lambda = 1$, the estimator $\widehat{f}_{n,m}$ reduces to the Bernstein estimator $\tilde{f}_{1,n,n}$.

4. Convergence

In this section, we derive some asymptotic properties of the proposed estimator $\widehat{f}_{n,m}$ when the sample size tends to infinity. First, we assume that λ and K are fixed. The following proposition gives the probability convergence of the proposed estimator $\widehat{f}_{n,m}$.

Proposition 1 (Probability convergence). *If $m = o(n/\log(n))$, then, for $x \in [0,1]$, we have*

$$\widehat{f}_{n,m}(x) \xrightarrow[n,m\to+\infty]{P} \lambda f(x) + (1-\lambda) f_2(x),$$

where $f_2(x) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, \sigma_j^2)(x)$, $\pi_j = \mathbb{E}(Z_{1j})$, $\mu_j = \mathbb{E}(X_1 \mid Z_{1j} = 1)$, $\sigma_j^2 = Var(X_1 \mid Z_{1j} = 1)$ for $j = 1, \ldots, K$ and $\xrightarrow{P}$ denotes the convergence in probability.

The proof of Proposition 1 necessitates the following technical Lemma.

Lemma 1. *Let $(S_n)_{n\geq 1}$ be a sequence of i.i.d. random variables in the space of square integral functions L^2 with a common mean μ and let $(T_n)_{n\geq 1}$ be a sequence of random variables. Hence,*

$$\mathbb{E}(\overline{S}_n | T_n) \xrightarrow[n\to+\infty]{L^2} \mu, \text{ where } \overline{S}_n = \frac{1}{n}\sum_{i=1}^{n} S_i,$$

where L^2 denotes the mean quadratic convergence L^2.

The proof of this lemma is reported in [35].

Proof of Proposition 1. First, using Lemma 1 and following the same steps as the proof of Theorem 4.4 in [35], we prove that $\widehat{\pi}_j \xrightarrow[n\to+\infty]{P} \pi_j$, $\lim_{n\to\infty} \widehat{\mu}_j \xrightarrow[n\to+\infty]{P} \mu_j$ and $\widehat{\sigma^2}_j \xrightarrow[n\to+\infty]{P} \sigma_j^2$. Then, according to Slutsky's Theorem, we obtain

$$\sum_{k=1}^{K} \widehat{\pi}_k \mathcal{N}(\widehat{\mu}_k, \widehat{\sigma}_k)(x) \xrightarrow[n\to+\infty]{P} \sum_{j=1}^{K} \pi_j \mathcal{N}(\mu_j, \sigma_j^2)(x). \tag{18}$$

Second, based on Theorem 3.1 in [16], we obtain

$$\tilde{f}_{1,n,m}(x) \xrightarrow[n\to+\infty]{P} f(x) \text{ for } x \in [0,1]. \tag{19}$$

In addition, referring to (18) and (19) and grounded on the application of Slutsky's Theorem, we conclude the proof. □

According to [21], the density $f(x)$ is a close approximation to the mixture density $f_2(x)$. Thus, the estimator $\widehat{f}_{n,m}(x)$ provides an approximation to the true density $f(x)$.

To study the asymptotic normality of the estimator $\widehat{f}_{n,m}$ given by (17), we set forward the following assumptions in [36].

(A1) For almost $x \in [0,1]$ and for all $i,j,h = 1\ldots,\nu$, the partial derivatives $\partial g/\partial\xi_i$, $\partial^2 g/\partial\xi_i\partial\xi_j$ and $\partial^3 g/\partial\xi_i\partial\xi_j\partial\xi_h$ of the density g exist and satisfy that $\left|\frac{\partial g(x|\theta)}{\partial\xi_i}\right|$, $\left|\frac{\partial^2 g(x|\theta)}{\partial\xi_i\xi_j}\right|$ and $\left|\frac{\partial^3 g(x|\theta)}{\partial\xi_i\xi_j\xi_h}\right|$ are bounded, respectively, by J_i, J_{ij} and J_{ijh}, where J_i and J_{ij} are integrable and J_{ijh}, satisfies

$$\int_0^1 J_{ijh}(x)g(x|\widehat{\theta})dx < \infty.$$

(A2) The Fisher information matrix $I(\theta)$ is positively defined at $\widehat{\theta}$.

Proposition 2 (Normality asymptotic)**.** *Under the regularity conditions* ***(A1)–(A2)****, if* $f(x) > 0$ *for all* $x \in [0,1]$, $2 \le m \le (n/\log n)$ *and* $\lim_{n,m\to\infty} n^{2/3}/m = 0$*, then, we obtain*

$$n^{1/2}m^{-1/4}\Big[\widehat{f}_{n,m}(x) - \lambda f(x) - (1-\lambda)f_2(x)\Big] \xrightarrow[n,m\to+\infty]{\mathcal{D}} \mathcal{N}\Big(0,\lambda^2\gamma(x)\Big),$$

where $\gamma(x) = f(x)(4\pi x(1-x))^{-1/2}$*, for* $x \in]0,1[$*, and* $\xrightarrow{\mathcal{D}}$ *denotes the convergence in distribution.*

Proof of Proposition 2. Using Theorem 3.2 in [16], we obtain

$$n^{1/2}m^{-1/4}(\tilde{f}_{1,n,m}(x) - f(x)) \xrightarrow[n,m\to+\infty]{\mathcal{D}} \mathcal{N}(0,\gamma(x)).$$

Thus,

$$n^{1/2}m^{-1/4}(\lambda\tilde{f}_{1,n,m}(x) - \lambda f(x)) \xrightarrow[n,m\to+\infty]{\mathcal{D}} \mathcal{N}(0,\lambda^2\gamma(x)).$$

According to Theorem 3.1 in [36], we obtain $\sqrt{n}(\widehat{\theta}-\theta) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}(0, I(\theta)^{-1})$. Using the delta method, we obtain

$$\sqrt{n}(\tilde{f}_{2,n}(x|\widehat{\theta}) - f_2(x|\theta)) \xrightarrow[n\to+\infty]{\mathcal{D}} \mathcal{N}\Big(0, Df_2(x|\theta)I(\theta)^{-1}Df_2(x|\theta)^T\Big),$$

where $Df_2(x|\theta)$ is the Jacobian matrix of $f_2(x|\theta) = f_2(x)$ and $\tilde{f}_{2,n}(x|\widehat{\theta}) = \tilde{f}_{2,n}(x)$. Since $m^{-1/4} \to 0$ if $m \to \infty$, then, using Slutsky's Theorem, we conclude the proof. □

The following corollary is a consequence of the previous proposition which gives an asymptotic confidence interval of the density f, for a risk $\alpha \in]0,1[$.

Corollary 1. *The* $100(1-\alpha)\%$ *asymptotic confidence interval of* $f(x)$ *is given by*

$$\left(\widehat{f}_{n,m}(x) \pm \frac{z_{1-\frac{\alpha}{2}}\lambda\sqrt{\gamma(x)}}{\sqrt{n}m^{-1/4}}\right),$$

where $z_{1-\frac{\alpha}{2}}$ *is the normal* $(1-\frac{\alpha}{2})$ *quantile.*

In the next section, we study the performance of the proposed estimator in estimating different distributions by comparing it to the performances of the Bernstein estimator and of the Gaussian kernel estimator.

5. Numerical Studies

5.1. Comparison Study

In this section, we investigate the performance of the proposed estimator given in (17), through estimating different densities by comparing it to the performance of the Bernstein estimator defined in (2), the standard Gaussian kernel estimator defined in (1) and the Guan's estimator defined in (3). We apply the Bernstein estimator when the sample is

concentrated on the interval $[0,1]$. For this purpose, we need to make some suitable transformations in the different cases that are listed as follows:

1. Let us suppose that X is concentrated on a finite support $[a,b]$; then, we work with the sample values $Y_1,\ldots,Y_n$, where $Y_i=(X_i-a)/(b-a)$.
2. For the density functions concentrated on $\mathbb{R}$, we can use the transformed sample $Y_i=1/2+\pi^{-1}\arctan(X_i)$, which transforms the range to the interval $(0,1)$.
3. For the support $\mathbb{R}_+$, we can use the transformed sample $Y_i=X_i/(1+X_i)$, which transforms the range to the interval $(0,1)$.

If the support is infinite, say, $\mathbb{R}=(-\infty,\infty)$, we can consider $[x_1,x_t]\subset[a,b]$ as the finite support of f, where x_1 and x_t are the minimum and the maximum order, respectively. We choose a and b such that $F(a)$ and $1-F(b)$ are of $O(n^{-1})$, $a<x_1$, and $b>x_t$, where F is the distribution function [27]. Then, we can use the transformed sample, which transforms $[x_1,x_t]$ to the interval $[0,1]$ mentioned in the case 1.

In the simulation study, three sample sizes were considered, $n=50$, $n=100$, and $n=200$, as well as the following density functions:

(a) The beta mixture density $0.5\mathcal{B}(3,9)+0.5\mathcal{B}(9,3)$;
(b) The beta mixture density $0.5\mathcal{B}(3,1)+0.5\mathcal{B}(10,10)$;
(c) The normal mixture density $1/4\mathcal{N}(2,1)+3/4\mathcal{N}(-3,1)$;
(d) The chi-squared $\chi_n(2)$ density.
(e) The gamma mixture density $0.5\mathcal{G}(1,6)+0.5\mathcal{G}(6,1)$;
(f) The gamma mixture density $0.5\mathcal{G}(1,2)+0.5\mathcal{G}(4,2)$.

Our sample was decomposed into a learning sample of a size of $2/3$ of the considered sample, on which the various statistical methods were constructed, and a second sample of a size of $1/3$ of the considered sample, on which the predictive performance of the three methods were tested. For each density function f and sample size n, we computed the integrated squared error (ISE), the integrated absolute error (IAE) and the Kullback–Leibler divergence (KL) of the estimator $\widehat{f}_{n,m}$ over $N=500$ trials.

$$\widehat{ISE}=\frac{1}{N}\sum_{k=1}^{N}ISE(\widehat{f}_k),\ \widehat{IAE}=\frac{1}{N}\sum_{k=1}^{N}IAE(\widehat{f}_k)\text{ and }\widehat{KL}=\frac{1}{N}\sum_{k=1}^{N}KL(\widehat{f}_k),$$

where $\widehat{f}_k$ is the estimator computed from the kth sample and

$$\begin{aligned}ISE[\widehat{f}_k] &= \int_0^1\left(\widehat{f}_k(x)-f(x)\right)^2dx,\ IAE(\widehat{f}_k)=\int_0^1|\widehat{f}_k(x)-f(x)|dx,\\ KL(\widehat{f}_k|f) &= \int_0^1\widehat{f}_k(x)\log\frac{\widehat{f}_k(x)}{f(x)}dx.\end{aligned}$$

Indeed, it is advised to consider a learning sample bigger than a testing sample. In this work, our sample was decomposed into a learning sample of a size of $2/3$ of the considered sample, on which the various statistical methods were constructed, and a second sample of a size of $1/3$ of the considered sample, on which the predictive performances of the three methods were tested. Each run of the proposed estimator performed the following steps:

- We first generated a random sample $(X_i)_{1\leq i\leq n}$ of size n from the models' density $(a)-(f)$.
- We then split the generated data into a training set of a size of $2/3$ of the considered sample and a test set of a size of $1/3$ of the considered sample.
- We applied the proposed estimator, using the observed data X_i only from the training set, in order to estimate the density function.
- The test set was then used to compute the estimation errors $\widehat{ISE}$, $\widehat{IAE}$ and $\widehat{KL}$.

To select the optimal parameter K, we used the Gap Statistics algorithm [37]. We considered a Monte Carlo experiment to select the optimal choice of the degree m of the Bernstein polynomial and the bandwidth h of the kernel estimator, for each point $x\in[0,1]$.

We determined the parameters m (for $1 \leq m \leq 300$) and h (for $h = i/1000$ with $1 \leq i \leq 300$), which minimized the ISE, which was approximated by the $\widehat{ISE}$.

We considered $N = 500$ random samples of sizes $n = 50$, $n = 100$ and $n = 200$.

Table 1. Average $\widehat{ISE}$ for $N = 500$ trials of Bernstein estimator, standard Gaussian kernel estimator and the proposed estimator $\widehat{f}_{n,m}$, for $n = 50$, $n = 100$ and $n = 200$. The bold values indicate the smallest values of ISE.

Density	n	Proposed Estimator	Bernstein Estimator	Kernel Estimator	Guan's Estimator
(a)	50	**0.092242**	0.096684	0.197497	0.140323
	100	**0.091364**	0.092242	0.174251	0.089129
	200	**0.075532**	0.079299	0.148143	0.086305
(b)	50	1.157827	1.215347	0.530446	**0.816906**
	100	**0.235402**	0.306704	0.482152	0.276573
	200	**0.199786**	0.289870	0.474805	0.255716
(c)	50	**0.001423**	1.808252	2.222369	1.035522
	100	**0.000384**	1.410606	1.641689	1.014602
	200	**0.000227**	1.348292	1.077352	0.994346
(d)	50	0.525192	2.812448	4.936701	0.589465
	100	**0.492752**	2.483141	2.331765	0.579595
	200	**0.162917**	0.898103	1.154646	0.507362
(e)	50	2.180849	2.231986	2.424656	**1.084340**
	100	2.050098	2.133496	2.295932	**0.835670**
	200	2.042379	2.086204	2.053453	**0.717715**
(f)	50	**0.313388**	0.896995	1.397111	0.663889
	100	**0.253988**	0.656400	0.762742	0.516530
	200	**0.186290**	0.577408	0.417980	0.472094

Table 2. Average $\widehat{IAE}$ for $N = 500$ trials of Bernstein estimator, standard Gaussian kernel estimator and the proposed estimator $\widehat{f}_{n,m}$, for $n = 50$, $n = 100$ and $n = 200$. The bold values indicate the smallest values of IAE.

Density	n	Proposed Estimator	Bernstein Estimator	Kernel Estimator	Guan's Estimator
(a)	50	**0.250241**	**0.250241**	0.391072	0.251641
	100	**0.196423**	0.207109	0.367361	0.232399
	200	**0.180536**	0.191673	0.348499	0.214117
(b)	50	0.855008	0.823416	**0.621137**	0.747562
	100	**0.417735**	0.457722	0.669027	0.438088
	200	**0.386280**	0.455983	0.583057	0.423595
(c)	50	**0.035161**	0.971720	1.238839	0.948451
	100	**0.019331**	0.960044	1.157838	0.944000
	200	**0.013233**	0.923259	0.953675	0.931293
(d)	50	0.669269	0.807667	1.942776	**0.648617**
	100	0.657815	1.486974	1.415036	**0.642986**
	200	**0.351205**	1.573886	0.881866	0.637942
(e)	50	0.830317	0.834203	1.499828	**0.678362**
	100	0.706770	0.713225	1.374609	**0.645112**
	200	0.681767	0.692357	1.225502	**0.620315**
(f)	50	**0.453948**	0.640596	1.068483	0.608421
	100	**0.414676**	0.714844	0.782939	0.584572
	200	**0.383248**	0.883320	0.516527	0.584320

Table 3. Average $\widehat{KL}$ for $N = 500$ trials of Bernstein estimator, standard Gaussian kernel estimator and the proposed estimator $\widehat{f}_{n,m}$, for $n = 50$, $n = 100$ and $n = 200$. The bold values indicate the smallest values of KL.

Density	n	Proposed Estimator	Bernstein Estimator	Kernel Estimator	Guan's Estimator
(a)	50	**0.025048**	0.025048	0.289818	0.066817
	100	**0.012086**	0.015023	0.081830	0.058541
	200	0.003256	**0.001788**	0.060468	0.029419
(b)	50	1.088053	1.120246	**0.575298**	0.667079
	100	**0.284795**	0.325933	0.381406	0.659105
	200	0.255697	0.310759	**0.150096**	0.654338
(c)	50	**2.689781**	3.871172	4.732324	3.360316
	100	**0.011844**	3.591156	4.196295	3.356093
	200	**0.009426**	3.505050	3.169852	3.318666
(d)	50	0.976450	0.870222	4.702359	**0.878995**
	100	0.960633	1.572251	3.031783	**0.763044**
	200	**0.281862**	1.584022	1.537355	0.742696
(e)	50	**1.549035**	1.560172	5.538142	1.799133
	100	**1.207084**	1.217043	3.498273	1.557894
	200	**1.153420**	1.169645	1.322299	1.387843
(f)	50	**0.528337**	1.052789	1.952294	0.651422
	100	**0.292017**	0.805962	1.191070	0.537625
	200	**0.062893**	0.589790	0.679154	0.339442

Departing from Tables 1–3 and Figure 1, we deduce the following:

- The results displayed in Tables 1–3 show that the $\widehat{ISE}$, $\widehat{IAE}$ and $\widehat{KL}$ decreased as the sample size increased.
- Using the proposed estimator, we obtained better results than those given by the other estimators in a large part of the cases.
- The figures 2 and 3 give a better sense of where the error is located.
- For the case (e) of the gamma mixture, the average $\widehat{ISE}$ and $\widehat{IAE}$ of Guan's estimator (1.3) were smaller than those obtained by the proposed density estimator (3.4) and the Bernstein estimator (1.2). However, in all the other cases, using an appropriate choice of the degree m, the average $\widehat{ISE}$ and $\widehat{IAE}$ of the proposed density estimator (3.4) were smaller than what achieved by the kernel estimator (1.1), the Bernstein estimator (1.2) and Guan's estimator (1.3), even when the sample size was large for same cases.
- When we changed the parameters of the gamma mixture density in the sense that we had a smaller bias, our estimator was more competitive than the other approaches and we obtained better results.
- Almost in all considered cases, the average $\widehat{KL}$ of the density estimator (17) was smaller than that obtained by the Bernstein estimator defined in (2), that of the kernel estimator defined in (1) and that of Guan's estimator.
- In the considered distribution $0.5\mathcal{B}(3,9) + 0.5\mathcal{B}(9,3)$, by choosing the appropriate m, the curve of the proposed distribution estimator (3.4) was closer to the true distribution than that of Guan's estimator (1.3), even when the sample size was very large.

Referring to Figures 2 and 3, we infer the following:

- None of the estimators for the gamma mixture density $0.5\mathcal{G}(6,1) + 0.5\mathcal{G}(1,6)$ had good approximations near $x = 0$. However, the $\widehat{ISE}$ of the proposed estimator was closer to zero than that of the Bernstein estimator and the kernel estimator, especially near the edge $x = 1$.
- Guan's estimator and the kernel estimator for the normal mixture density $0.25\mathcal{N}(2,1) + 0.75\mathcal{N}(-3,1)$ had good approximations near $x = 0$. However, the $\widehat{ISE}$ of the pro-

posed estimator was closer to zero than that of the other estimators, especially near the two edges.

Therefore, we note that, for difficult distributions that diverge at the boundaries, the proposed method would fail, but not as badly as the standard methods without shrinkage. In addition, the performed simulations revealed that, on average, the proposed approach could lead to satisfactory estimates near the boundaries, better than the classical Bernstein estimator.

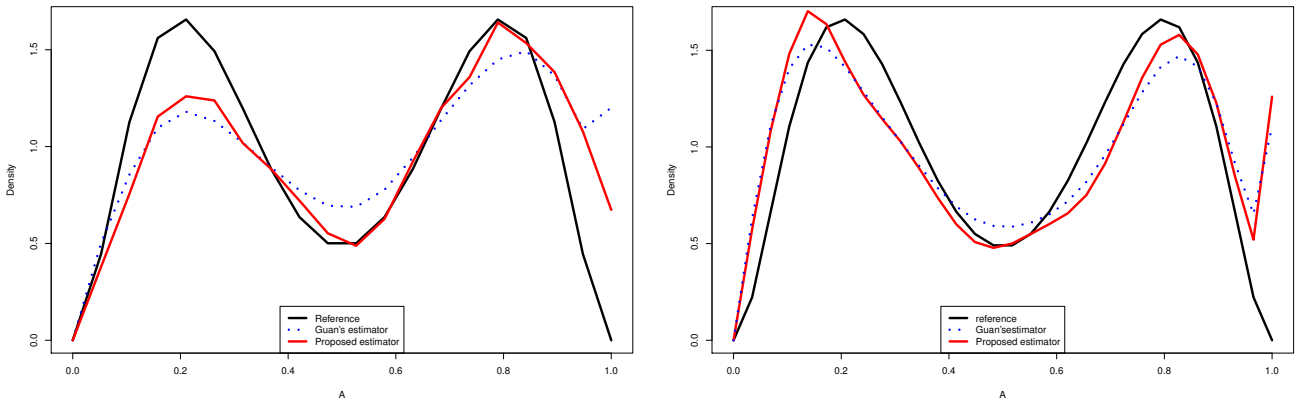

Figure 1. Quantitative comparison between the proposed estimator and Guan's estimator of $0.5\mathcal{B}(3,9) + 0.5\mathcal{B}(9,3)$ for $n = 50$ (**left**) and $n = 100$ (**right**).

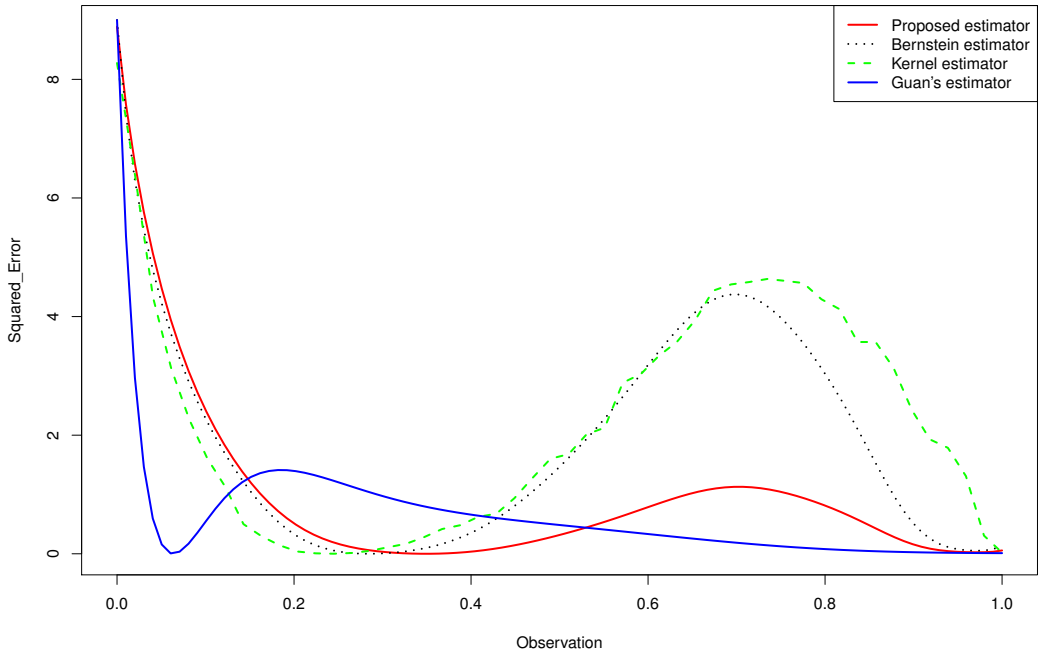

Figure 2. Quantitative comparison among the mean squared error of the kernel estimator, the Bernstein estimator, the Guan's estimator and the proposed estimator of $0.5\mathcal{G}(1,6) + 0.5\mathcal{G}(6,1)$ for $n = 200$.

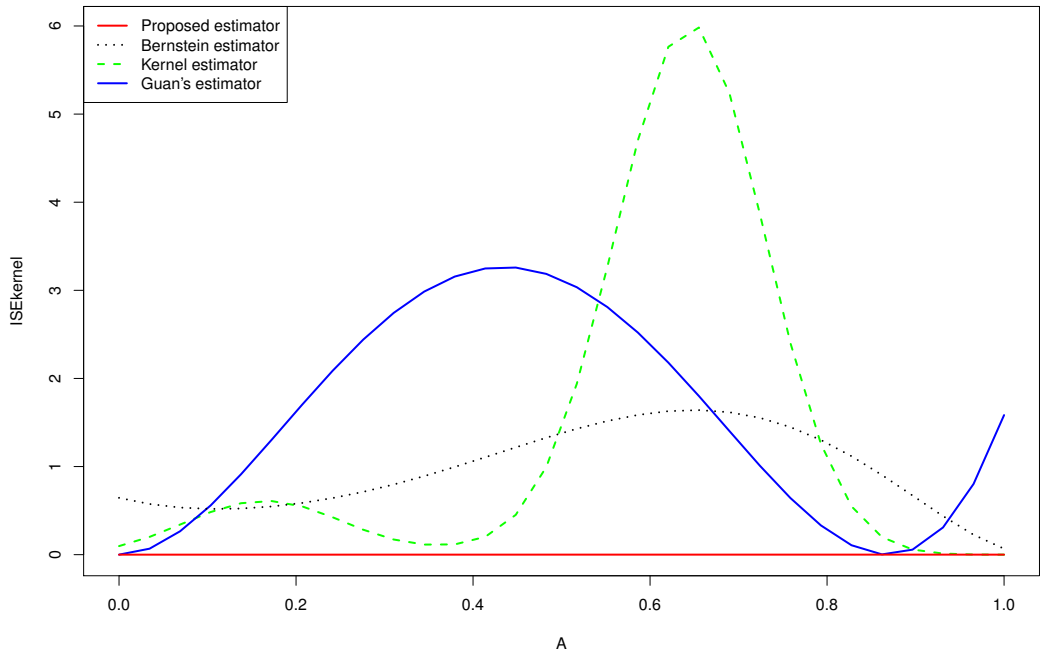

Figure 3. Quantitative comparison among the mean squared error of the kernel estimator, the Bernstein estimator, the Guan's estimator and the proposed estimator of $0.25\mathcal{N}(2,1) + 0.75\mathcal{N}(-3,1)$ for $n = 200$.

5.2. Real Dataset

5.2.1. COVID-19 Data

In this subsection, we consider the COVID-19 data displayed in the INED website https://dc-covid.site.ined.fr/fr/donnees/france/ (accessed on 16 February 2022). These data concern the numbers of deaths due to COVID-19 in France (daily) from 21 March 2021, for 454 days. These data are such that $\min_i(x_i) = 605$ and $\max_i(x_i) = 0$. Then, it is convenient to assume that the density of the numbers of deaths is defined on the interval $[0, 605]$ and transform the data into the interval unit. The Monte Carlo procedure was performed and resulted in $h = 0.07659612$ for the standard kernel estimator defined in (1.1), $m_1 = 20$ for the Bernstein estimator defined in (1.2), the proposed estimator, and $m_2 = 12$ for Guan's estimator. These estimators are exhibited in Figure 1 (right panel) along with a histogram of the data. All the estimators are smooth and seem to capture the pattern highlighted by the histogram. We record that the proposed estimator outperformed the other estimators near the boundaries.

5.2.2. Tuna Data

The last example concerns the tuna data reported in [38]. The data are derived from an aerial line transect survey of Southern Bluefin Tuna in the Great Australian Bight. An aircraft with two spotters on board flew randomly over allocated line transects. These data correspond to the perpendicular sighting distances (in miles) of 64 detected tuna schools to the transect lines. The survey was conducted in summer when tuna data tend to stay on the surface. The data are such that $\min_i(x_i) = 0.19$ and $\max_i(x_i) = 16.26$. The Monte Carlo procedure was performed and resulted in $h = 0.1079$ for the standard kernel estimator defined in (1), $m_1 = 13$ for the Bernstein estimator defined in (2) and the proposed estimator, and $m_2 = 6$ for Guan's estimator. These estimators are illustrated in Figure 4 (left panel) along with a histogram of the data. All the estimators are smooth and seem to capture the pattern highlighted by the histogram. We assert that the proposed estimator outperformed the other estimator, especially near the boundaries.

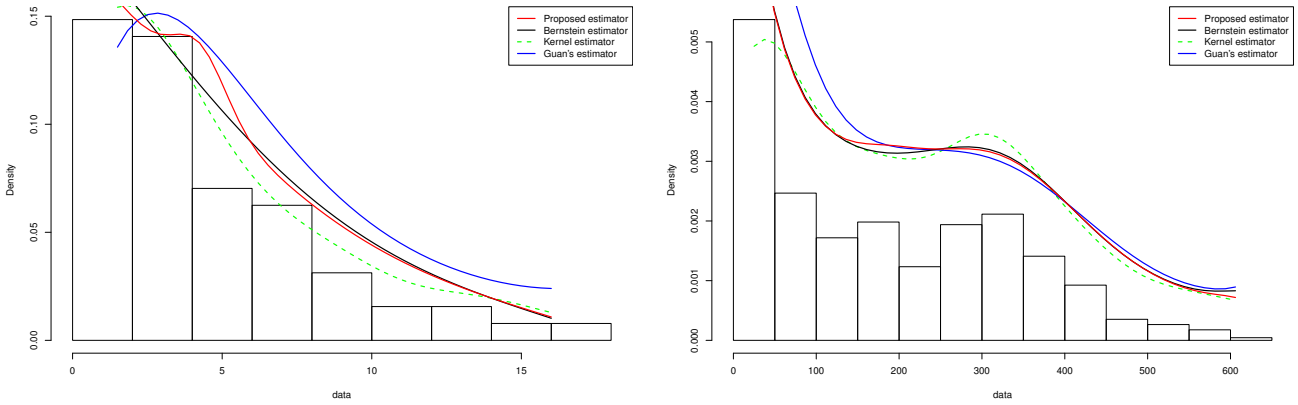

Figure 4. Qualitative comparison among the kernel estimator defined in (1), the Bernstein estimator defined in (2), Guan's estimator (3) and the proposed density estimator (17) of Tuna data (**left**) and of COVID-19 data (**right**).

6. Conclusions

In this paper, we propose a shrinkage estimator of a density function based on the Bernstein density estimator and using a finite Gaussian mixture density. This method rests on three steps. The first step consists of considering the Bernstein estimator $\tilde{f}_{1,n,m}$. The second relies upon the Gaussian Mixture density $\tilde{f}_{2,n}$ as an estimator of the unknown density f. The last step consists of considering the shrinkage form $\lambda\tilde{f}_{1,n,m} + (1-\lambda)\tilde{f}_{2,n}$ and EM algorithm in order to estimate the parameter λ. The asymptotic properties of this estimator were established. Afterwards, we demonstrate the effectiveness of the proposed method using some simulated and real data. We clarify how it can lead to very satisfactory estimates near the boundaries and in terms of ISE, IAE and KL. Eventually, we would simply assert that our research work is a step that may be taken further, extended and built upon as it lays the ground and paves the way for future works to elaborate a semi-parametric regression estimator using the shrinkage method. We also plan to work on the case where λ is a random variable. Another future research direction would be to extend our findings to the setting of serially dependent observations.

Author Contributions: Conceptualization, A.M. and Y.S.; Data curation, S.H.; Investigation, S.H. and Y.S.; Methodology, S.H., A.M. and Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research study received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Stat.* **1956**, *27*, 832–837. [CrossRef]
2. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076. [CrossRef]
3. Härdle, W. *Smoothing Techniques with Implementation in S*; S. Springer Science and Business Media: Berlin, Germany, 1991.
4. Schuster, E.F. Incorporating support constraints into nonparametric estimators of densities. *Comm. Stat. Theory Methods* **1985**, *14*, 1123–1136. [CrossRef]
5. Müller, H.-G. Smooth optimum kernel estimators near endpoints. *Biometrika* **1991**, *78*, 521–530. [CrossRef]
6. Müller, H.-G. On the boundary kernel method for nonparametric curve estimation near endpoints. *Scand. J. Statist.* **1993**, *20*, 313–328.
7. Müller, H.-G.; Wang, J.-L. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrika* **1994**, *50*, 61–76. [CrossRef]

8. Lejeune, M.; Sarda, P. Smooth estimators of distribution and density functions. *Comput. Stat. Data Anal.* **1992**, *14*, 457–471. [CrossRef]
9. Jones, M.C. Simple boundary correction for density estimation kernel. *Stat. Comput.* **1993**, *13*, 135–146. [CrossRef]
10. Chen, S.X. Beta kernel estimators for density functions. *Comput. Stat. Data Anal.* **1999**, *31*, 131–145. [CrossRef]
11. Chen, S.X. Probability density function estimation using gamma kernels. *Ann. Inst. Stat. Math.* **2000**, *52*, 471–480. [CrossRef]
12. Leblanc, A. A bias-reduced approach to density estimation using Bernstein polynomials. *J. Nonparametr. Stat.* **2010**, *22*, 459–475. [CrossRef]
13. Slaoui Y. Bias reduction in kernel density estimation. *J. Nonparametr. Stat.* **2018**, *30*, 505–522. [CrossRef]
14. Vitale,R.A. A bernstein polynomial approach to density function estimation. *Stat. Inference Relat. Topics* **1975**, *2*, 87–99.
15. Ghosal, S. Convergence rates for density estimation with Bernstein polynomials. *Ann. Stat.* **2000**, *29*, 1264–1280. [CrossRef]
16. Babu, G.J.; Canty, A.J.; Chaubey, Y.P. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *J. Stat. Plan. Inference* **2002**, *105*, 377–392. [CrossRef]
17. Kakizawa, Y. Bernstein polynomial probability density estimation. *J. Nonparametr. Stat.* **2004**, *16*, 709–729. [CrossRef]
18. Rao, B.L.S.P. Estimation of distribution and density functions by generalized Bernstein polynomials. *Indian J. Pure Appl. Math.* **2005**, *36*, 63–88.
19. Igarashi, G.; Kakizawa, Y. On improving convergence rate of Bernstein polynomial density estimator. *J. Nonparametr. Stat.* **2014**, *26*, 61–84. [CrossRef]
20. Slaoui, Y.; Jmaei, A. Recursive density estimators based on Robbins-Monro's scheme and using Bernstein polynomials. *Stat. Interface* **2019**, *12*, 439–455. [CrossRef]
21. Li, J.Q.; Barron, A.R. Mixture density estimation. *Adv. Neural Inf. Process. Syst.* **2000**, *12*, 279–285 .
22. Pearson, K. Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **1984**, *185*, 71–110.
23. McLachlan, G.; Peel, D. *Finite Mixture Models*; John Wiley and Sons: New York, NY, USA 2004.
24. Roeder, K.; Wasserman, L. Practical Bayesian density estimation using mixtures of normals. *J. Am. Stat. Assoc.* **1997**, *92*, 894–902. [CrossRef]
25. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
26. Leroux, B. Consistent estimation of a mixing distribution. *Ann. Stat.* **1992**, *20*, 1350–1360. [CrossRef]
27. Guan, Z. Efficient and robust density estimation using Bernstein type polynomials. *J. Nonparametr. Stat.* **2016**, *28*, 250–271. [CrossRef]
28. James, W.; Stein, C. Estimation with quadratic loss. In *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1992; pp. 443–460.
29. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1997**, *39*, 1–22.
30. Wu, C.J. On the convergence properties of the EM algorithm. *Ann. Stat.* **1983**, *11*, 95–103. [CrossRef]
31. Stein, C. Estimation of the mean of a multivariate normal distribution. *Ann. Stat.* **1981**, *9*, 1135–1151. [CrossRef]
32. Oman, S.D. Contracting towards subspaces when estimating the mean of a multivariate normal distribution. *J. Multivar. Anal.* **1982**, *12*, 270–290. [CrossRef]
33. Oman, S.D. Shrinking towards subspaces in multiple linear regression. *Technometrics* **1982**, *24*, 307–311. [CrossRef]
34. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2006.
35. Zitouni, M.; Zribi, M.; Masmoudi, A. Asymptotic properties of the estimator for a finite mixture of exponential dispersion models. *Filomat* **2018**, *32*, 6575–6598. [CrossRef]
36. Redner, R.A.; Walker, H.F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **1984**, *26*, 195–239. [CrossRef]
37. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser.* **2001**, *63*, 411–423. [CrossRef]
38. Chen, S.X. Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **1996**, *83*, 329–341. [CrossRef]