ON THE SMOOTHED PARAMETRIC ESTIMATION OF MIXING PROPORTION UNDER FIXED DESIGN REGRESSION MODEL

Y. S. Ramakrishnaiah¹, Manish Trivedi², Konda Satish³

ABSTRACT

The present paper revisits an estimator proposed by Boes (1966) – James (1978), herein called BJ estimator, which was constructed for estimating mixing proportion in a mixed model based on independent and identically distributed (i.i.d.) random samples, and also proposes a completely new (smoothed) estimator for mixing proportion based on independent and not identically distributed (non-i.i.d.) random samples. The proposed estimator is nonparametric in true sense based on known “kernel function” as described in the introduction. We investigated the following results of the smoothed estimator under the non-i.i.d. set-up such as (a) its small sample behaviour is compared with the unsmoothed version (BJ estimator) based on their mean square errors by using Monte-Carlo simulation, and established the percentage gain in precision of smoothed estimator over its unsmoothed version measured in terms of their mean square error, (b) its large sample properties such as almost surely (a.s.) convergence and asymptotic normality of these estimators are established in the present work. These results are completely new in the literature not only under the case of i.i.d., but also generalises to non-i.i.d. set-up.

Key words: mixture of distributions, mixing proportion, smoothed parametric estimation, fixed design regression model, mean square error, optimal band width, strong consistency, asymptotic normality.

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and not identically distributed (non-i.i.d.) random variables with continuous distribution functions (d.f.s) $\{F_i(x), 1 \leq i \leq n\}$. Let $H(x)$ be a continuous cumulative distribution function (cdf) of mixture of component cdfs $H_1(x), \ldots, H_m(x)$, ($m \geq 2$) such that $H(x) = \sum_{j=1}^{m} p_j H_j(x)$, where $\{p_j; 1 \leq j \leq m\}$ is a set of mixing proportions satisfying (i) $0 < p_j < 1$, (ii) $\sum_{j=1}^{m} p_j = 1$. Let $\tilde{H}_n(x) = n^{-1} \sum_{i=1}^{n} F_i(x) \to H(x)$ as $n \to \infty$ and $\bar{H}_j(x) = n_j^{-1} \sum_{i=1}^{n_j} F_{ji}(x) \to H_j(x)$.

¹ Faculty of Statistics, Osmania University in Hyderabad, India. E-mail: ysrkou@gmail.com.
² Faculty of Statistics, School of Sciences, Indira Gandhi National Open University in New Delhi, India. E-mail: Manish_trivedi@ignou.ac.in. ORCID ID: https://orcid.org/0000-0002-5790-6546.
³ Faculty of Statistics, Aurora College in Hyderabad, India. E-mail: statishlaks@gmail.com. ORCID ID: https://orcid.org/0000-0003-4119-8773.
\(n_j \to \infty, \ j=1,2,\ldots,m; \ H(x), H_j(x)\) are known d.f.s. The problem of estimation of mixing proportions \(p_j\) in a mixture

\[
H(x) = p_1 H_1(x) + p_2 H_2(x) + \ldots + p_m H_m(x) \tag{1.1}
\]

of \(m\) known distributions \(H_j(x)\) is investigated based on independent random samples of sizes \(n, n_j\) generated from the fixed design regression models

\[
X_i = \beta t_i + \epsilon_i, \ 1 \leq i \leq n, \ \epsilon_i \sim i.i.d. F(x) \tag{1.2}
\]

\[
X_{ji} = \beta_j t_{ji} + \epsilon_{ji}, \ 1 \leq i \leq n_j, \ j = 1, 2,\ldots,m, \ \epsilon_{ij} \sim i.i.d. F_j(x), \tag{1.3}
\]

\(\beta\)'s and \(t\)'s are known reals satisfying the model conditions

\[
\beta_j > 0, \ \sum_{i=1}^{n} t_i = 0 \text{ and } \frac{1}{n} \sum t_i^2 = o(n^{-1}). \tag{1.4}
\]

Note that \(t_i = \mp \frac{i}{n^{\frac{1}{2}}}i = \mp1, \mp2,\ldots, \mp n, \ \delta \geq \frac{3}{2} \text{ fulfill (1.4) and } H_n(x) = n^{-1} \sum_{i=1}^{n} F_i(x) = F(x) + O(\frac{1}{n} \sum t_i^2) + \ldots = H(x) + o(n^{-1}) \text{ and } H_{nj}(x) = n_j^{-1} \sum_{i=1}^{n_j} F_{ji}(x) \to H_j(x), \ j=1,2 \text{ as } n_j \to \infty.

Mixture distributions have been used in a wide variety of numerous applications in such diversified fields as physics, chemistry, biology, social sciences and others. Many typical problems in which such mixtures occur have been well described in a series of research papers. Karl Pearson (1894) dealt with the application of normal mixtures to the theory of evolution, which considered the first paper in the mixtures of distributions. Acheson and McElwee (1951), who identified failures in an electronic tube in gaseous defects, mechanical defects and normal deterioration of the cathode. One can find the proportion of the population which will fail in each cause to redesign the system or to improve the methods of manufacturing process. Apart from this, it would be desirable to know the distribution of defectives for each cause. Mendenhall and Hader (1958) studied censored life testing as a mixed failure populations. They suggested an example that the engineer may identify the product as defective/failure and nondefective by two or more different types of causes. Hosmer (1973) studied characteristics such as sex, age, and length of halibut (fish). Odell and Basu (1976) applied them in the field of remote sensing to estimate the crop acreages from remote sensors on orbiting satellites.

We shall show some of the typical problems which were described in Choi and Bulgren (1968):

1. In fishery biology, it is often desired to measure certain characteristics in a natural population of fish. For this purpose samples of fish are taken and the desired trait is measured for each fish in the sample. However, many characteristics vary markedly with the age of the fish. Then, the trait has a distinct distribution for each age group so that the population has a mixture of distributions.

2. A geneticist analyses the inheritance of qualitative characters. In general, such characters vary continuously over some intervals of real numbers so that a given genotype may be able to produce phenotypic values over an interval of real numbers. Then, the phenotypic value which the geneticist observes has a mixture of distributions, each of which is given by a genotype.
3. In photographing the absorption spectrum of an ionized atom, we obtain a photograph of a constantly varying intensity distribution on the photographic plate, and not a series of discrete "lines". This phenomenon is caused by several effects (such as the Doppler effect) and it is accepted in spectroscopy that an intensity distribution whose graph can be approximated closely by that of a normal density function belongs to every theoretical "line". Then the graph of the whole spectrum section can be considered as a mixture of normal density functions.

Other references to mixed failure populations are given in papers by Davis (1952), Epstein (1953), Herd (1953), Steen and Wilde (1952), Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1988), Lindsay (1995) and McLachlan and Peel (2000), Fu (1968- Pattern Recognition), Varli et al. (1975- Pattern Recognition), Clark (1976- Geology), Macdonald and Pitcher (1979- Fisheries), Bruni et al. (1985- Genetics), Merz (1980- Physics) and Christensen et al. (1980- Nuclear Physics).

The applications of finite mixture distributions describing mixture populations for non-i.i.d. sequence of variables are given below:

<table>
<thead>
<tr>
<th>Area</th>
<th>Characteristic $X_i$</th>
<th>Distribution function $F_i(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survival Analysis</td>
<td>life time of components produced by $i^{th}$ machine operated by $i^{th}$ foreman.</td>
<td>Life time distribution of various products</td>
</tr>
<tr>
<td>Nutritional Studies</td>
<td>weight for age/height for age/weight for height of $i^{th}$ infant of $i^{th}$ origin or group.</td>
<td>Distribution of weight for age/height of $i^{th}$ infant</td>
</tr>
<tr>
<td>Fisheries</td>
<td>Fish length or weight of age of $i^{th}$ fish.</td>
<td>Distribution of weight/length of $i^{th}$ fish</td>
</tr>
</tbody>
</table>

The mixing model with two component populations becomes

$$H(x) = pH_1(x) + (1-p)H_2(x).$$

Here, $X_i$ is the characteristic with distribution function $F_i(x)$ assuming

$$\bar{F}_n(x) = n^{-1}\sum_{i=1}^{n} F_i(x) = \bar{H}_n(x) \rightarrow H(x)$$

$$\bar{F}_{jn}(x) = n^{-1}\sum_{i=1}^{n} F_{ji}(x) = \bar{H}_{jn}(x) \rightarrow H_j(x) \text{ as } n \rightarrow \infty$$

and $n$, $n_1$ and $n_2$ are independent random sample sizes from mixed and component populations selected in such a way that $n = n_1 + n_2$ with $n_1 = \lfloor pn \rfloor$, $n_2 = \lfloor (1-p)n \rfloor$, $0 < p < 1$.

More details on such examples can be found in Choi and Bulgreen (1968), Harris (1958), Blischke (1965), Fu (1968- Pattern Recognition), Varli et al. (1975- Pattern Recognition), Clark (1976- Geology), Macdonald and Pitcher (1979- Fisheries), Odell and Basu (1976- Remote sensing), Bruni et al. (1985- Genetics), Merz (1980- Physics) and Christensen et al. (1980- Nuclear Physics), etc.
i.i.d case: The mixing model for i.i.d. case is
\[ F(x) = pG_1(x) + (1-p)G_2(x) \]  
(1.5)
where \( F(x), \, G_j(x); \, j=1,2 \) are cdfs of mixed and component populations respectively. The following estimator is studied in the literature.

**Boes (1966)-James (1978) (BJ) estimator:** Let \( F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \) be the empirical distribution function of a random sample \( X_i, 1 \leq i \leq n \) from a mixture of two known component distribution functions \( G_j, j=1,2 \). Boes (1966) proposed an estimator of \( p \) given by
\[ p_{n,1}(x) = \frac{\hat{F}_n(x) - G_2(x)}{G_1(x) - G_2(x)} \]  
(1.6)
and shown as a minimax unbiased estimator, and derived the Cramer-Rao lower bound. James (1978) considered the problem of estimating the mixing proportion in a mixture of two known normal distributions. He studied the simple estimators based on (a) the number of observations lower than a fixed point \( r \), (b) the numbers lower than \( s \) and greater than \( t \), and (c) the sample mean. Van Houwelingen (1974) used Boes (1966) estimator to estimate the mixing proportion by using frequency densities and obtained the Cramer-Rao lower bound. Jayalakshmi (2002) used BJ estimator using kernel-based empirical distribution and established that smoothing improves efficiency when the component distributions are known.

In the present work, we extend the idea of estimation of mixing proportion \( p \) in two directions:
- The estimators based on kernel based empirical d.f. called smoothed estimations are proposed under regression models (1.2) - (1.3).
- The proposed parametric estimators are based on independent, but not identically distributed (non-i.i.d.) samples generated by the fixed design regression models described by (1.2)-(1.3).

The main object of the present paper is to confine attention to \( m=2 \) case in the model (1.1) and to construct parametric estimators when component distributions are known, based on the usual empirical and kernel-based distribution functions defined by
\[ \hat{H}_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \leq x), \quad \tilde{H}_n(x) = n^{-1} \sum_{i=1}^{n} K\left(\frac{x-X_i}{a_n}\right), \]  
(1.7)
\( \{a_n\} \) being the smoothing sequence satisfying \( 0 < a_n \rightarrow 0, n a_n \rightarrow \infty \) defined by
\[ p_{n,1}(x) = \frac{\hat{H}_n(x) - H_2(x)}{H_1(x) - H_2(x)}, \quad p_{n,2}(x) = \frac{\tilde{H}_n(x) - H_2(x)}{H_1(x) - H_2(x)}. \]  
(1.8)

We study the small and large sample behaviour of the proposed parametric estimators and establish the superiority of smoothed estimator \( p_{n,2}(x) \) over unsmoothed one \( p_{n,1}(x) \) in the sense of minimum mean square error. The results of the present investigations for the non-i.i.d. sequences are completely new in the literature.

In section 2, we obtain the exact expressions for MSEs of the proposed estimators in order to establish the superiority of \( p_{n,2}(x) \) over \( p_{n,1}(x) \) for some fixed \( x \). Furthermore, large sample behaviour, such as asymptotic normality and rates
of a.s. convergence of the proposed estimators is also established. In section 3, the crucial choice of smoothing parameter ‘\(a_n\)’ in kernel based estimator \(p_{n,2}(x)\) is discussed and its value is determined by employing minimum mean square criterion. Section 4 deals with establishing superiority of \(p_{n,2}(x)\) over \(p_{n,1}(x)\). Section 5 explains the small sample comparisons by Monte Carlo method based on the samples generated by regression models.

2. Asymptotics of \(p_{n,1}(x)\) and \(p_{n,2}(x)\)

We now present the properties of both estimators \(p_{n,1}(x)\) and \(p_{n,2}(x)\) under the fixed design regression model (1.3). The properties such as mean square errors (MSEs), rates of a.s. convergence, and asymptotic normality of \(p_{n,1}(x)\) and \(p_{n,2}(x)\) are established. We first consider the representations of the proposed estimators: from the mixing model,

\[
H(x) = pH_1(x) + (1-p)H_2(x)
\]

\[
H(x) - H_2(x) = d_{12}(x)p \quad \text{for each } x, \text{ where } d_{12}(x) = H_1(x) - H_2(x)
\]

**A.S. Representations to \(p_{n,1}(x)\), \(p_{n,2}(x)\):** from (1.8) and (2.1),

\[
d_{12}(x)[p_{n,1}(x) - p] = \tilde{H}_n(x) - H(x) = n^{-1} \sum_{i=1}^{n} (I(X_i \leq x) - F_i(x) + F_i(x) - H(x))
\]

\[
= n^{-1} \sum_{i=1}^{n} Z_{1i}(x) + \tau_{n1}(x)
\]  \hspace{1cm} (2.2)

where by (1.4) and by the assumption on regression model

\[
\tau_{n1}(x) = n^{-1} \sum_{i=1}^{n} F_i(x) - H(x) = o(n^{-1} \sum_{i=1}^{n} t_i^2) = o(n^{-1})
\]  \hspace{1cm} (2.3)

Similarly, from (1.8) and (2.1)

\[
d_{12}(x)[p_{n,2}(x) - p] = \tilde{H}_n(x) - H(x) = \tilde{H}_n(x) - E \tilde{H}_n(x) + E \tilde{H}_n(x) - H(x)
\]

\[
= n^{-1} \sum_{i=1}^{n} \left[ K\left(\frac{x - X_i}{a_n}\right) - E K\left(\frac{x - X_i}{a_n}\right) \right] + \tau_{n2}(x)
\]

\[
= n^{-1} \sum_{i=1}^{n} Z_{2i}(x) + \tau_{n2}(x)
\]  \hspace{1cm} (2.4)

with \(\tau_{n2}(x) = o(n^{-1})\) as in (2.3).

2.1. Mean Square Error of \(p_{n,1}(x)\) and \(p_{n,2}(x)\)

We first consider the small sample property, i.e. MSEs of both estimators \(p_{n,j}(x)\), \(j=1,2\) for each \(x\) in the following, which will help in establishing the
superiority of smoothed estimator over unsmoothed version based on a random sample of exact size \( n \) under the regression model.

**Theorem 2.1.** Let \( \{X_i: 1 \leq i \leq n\} \) be a sequence of non-i.i.d. random variables with corresponding sequence of uniformly continuous distribution functions \( \{F_i(x): 1 \leq i \leq n\} \). If \( \{F_i\} \), the kernel density function \( k \) and \( \{a_n\} \) in (1.7) satisfy

\[ \begin{align*}
\text{AI:} & \quad \text{i) } F_i(x) \text{ is uniformly continuous distribution function with finite } q^{th} \text{ derivatives } F_i^{(q)}(x) < \infty, 1 \leq i \leq n \text{ and } \\
& \quad \text{ii) } \bar{H}_n(x) = n^{-1} \sum_{i=1}^{n} F_i(x) \to H(x) \text{ as } n \to \infty
\end{align*} \]

\[ \begin{align*}
\text{All:} & \quad \text{i) The kernel function satisfies } \mu_{2j}(K) = \int_{-\infty}^{\infty} t^{2j} dK(t) \neq 0 \text{ and } \\
& \quad \text{ii) } \psi_j(K) = 2 \int_{-\infty}^{\infty} t^{j} K(t) dK(t) < \infty, j = 0, 1, 2, 3, 4
\end{align*} \]

\[ \begin{align*}
\text{AllI:} & \quad \text{A sequence of bandwidths such that } \\
& \quad \text{i) } 0 < a_n \downarrow 0; na_n \to \infty \text{ as } n \to \infty \\
& \quad \text{ii) } na_n^4 \to 0 \text{ as } n \to \infty
\end{align*} \]

then

\[
\text{MSE } [p_{n,2}(x)] = \text{MSE}(p_{n,1}(x)) - d_{12}^{-2}(x)[ \frac{a_n}{n} \bar{H}_n^{(1)}(x) \psi_1(K) - \frac{a_n^4}{4} \bar{H}_n^{(2)}(x) \mu_2^2(K)] + O(a_n^2) + o(a_n^4)
\]

**Proof:** From (2.4),

\[
d_{12}^{-2}(x)n\text{MSE}[p_{n,2}(x)] = \text{Var}(n^{-1/2} \sum_{i=1}^{n} Z_{2i}(x)) + n \tau_{n2}^2
\]

where

\[
\begin{align*}
\sigma_{n2}^2 &= \text{Var}(n^{-1/2} \sum_{i=1}^{n} Z_{2i}(x)) \\
&= n^{-1} \sum_{i=1}^{n} E Z_{2i}^2(x) \\
&= n^{-1} \sum_{i=1}^{n} \sigma_{2i}^2 \\
&= E K^2(\frac{x-x_i}{a_n}) - E^2 K(\frac{x-x_i}{a_n}) \\
&= I_{1i} - I_{2i} \text{ (say)}
\end{align*}
\]

where

\[
\begin{align*}
I_{1i} &= E K^2(\frac{x-x_i}{a_n}) = \int K^2(\frac{x-u}{a_n}) dF_i(u) = \int F_i(x-a_n t) dK^2(t) \\
&= F_i(x)[dK^2(t) - F_i^{(1)}(x)a_n \int dK^2(t) + \frac{a_n^2}{2!} F_i^{(2)}(x) \int t^2 dK^2(t) - \frac{a_n^3}{3!} F_i^{(3)}(x) \int t^3 dK^2(t) \\
&\quad + \frac{a_n^4}{4!} F_i^{(4)}(x) \int t^4 dK^2(t) + o(a_n^4)]
\end{align*}
\]
\[ F_i(x) = F_i(1)(x) \psi_1(K) + \frac{a^2_n}{2!} F_i(2)(x) \psi_2(K) - \frac{a^3_n}{3!} F_i(3)(x) \psi_3(K) \]
\[ + \frac{a^4_n}{4!} F_i(4)(x) \psi_4(K) + o(a^6_n) \]  
(2.6)

where \( \psi_0(K) = 2 \int_{-\infty}^{\infty} K(t) dK(t) = 2 \int_{0}^{1} \gamma dy = 1 \), while

\[ I_{2i} = E K \left( \frac{x - X_i}{a_n} \right) = \int F_i(x) - a_n t dK(t) \]
\[ = F_i(x) + \frac{a^2_n}{2!} F_i(2)(x) \int_{-\infty}^{\infty} t^2 dK(t) + \frac{a^3_n}{3!} F_i(3)(x) \int_{-\infty}^{\infty} t^4 dK(t) + o(a^4_n) \]
\[ =: F_i(x) + \frac{a^2_n}{2} F_i(2)(x) \mu_2(K) + \frac{a^4_n}{4!} F_i(4)(x) \mu_4(K) + o(a^5_n) \]  
(2.7)

From (2.5) and (2.7),

\[ \sigma^2_{n2} = n^{-1} \sum_{i=1}^{n} \left[ \left( F_i(x) - a_n F_i(1)(x) \psi_1(K) + \frac{a^2_n}{2} F_i(2)(x) \psi_2(K) - \frac{a^3_n}{3!} F_i(3)(x) \psi_3(K) \right. \right. \]
\[ + \frac{a^4_n}{4!} F_i(4)(x) \psi_4(K) + o(a^6_n) \]
\[ \left. \left. - (F_i(x) + \frac{a^2_n}{2} F_i(2)(x) \mu_2(K) + \frac{a^4_n}{4!} F_i(4)(x) \mu_4(K) + o(a^5_n))^2 \right] \right] \]
\[ = n^{-1} \sum_{i=1}^{n} F_i(x)(1 - F_i(x)) - a_n \tilde{H}_n^{(1)}(x) \psi_1(K) + \frac{a^2_n}{2} \tilde{H}_n^{(2)}(x) \psi_2(K) \]
\[ \frac{a^3_n}{3!} \tilde{H}_n^{(3)}(x) \psi_3(K) - \frac{a^4_n}{4!} \tilde{H}_n^{(4)}(x) \psi_4(K) \]
\[ - \frac{a^3_n}{3} n^{-1} \sum_{i=1}^{n} F_i(x) F_i(2)(x) \mu_2(K) \]
\[ - \frac{a^4_n}{4!} n^{-1} \sum_{i=1}^{n} F_i(x) F_i(4)(x) \mu_4(K) + o(a^5_n) \]
\[ = n^{-1} \sum_{i=1}^{n} F_i(x)(1 - F_i(x)) - a_n \tilde{H}_n^{(1)}(x) \psi_1(K) \]
\[ + a^2_n \tilde{H}_n^{(2)}(x) \psi_2(K) - n^{-1} \sum_{i=1}^{n} F_i(x) F_i(2)(x) \mu_2(K) + O(a^5_n) \]  
(2.8)

and from (2.7),

\[ \tau^2_{n2} = (n^{-1} \sum_{i=1}^{n} E K \left( \frac{x - X_i}{a_n} \right) - H(x))^2 \]
\[ = \left[ \tilde{H}_n(x) - H(x) + \frac{a^2_n}{2} \tilde{H}_n^{(2)}(x) \mu_2(K) + \frac{a^4_n}{4!} \tilde{H}_n^{(4)}(x) \mu_4(K) + o(a^6_n) \right]^2 \]
\[ =: \left[ \xi_{n,0}(x) + a^2_n \xi_{n,2}(x) + \frac{a^4_n}{4!} \xi_{n,4}(x) + o(a^6_n) \right]^2 \]
\[ = \xi_{n,0}(x) [\xi_{n,0}(x) + 2a^2_n \xi_{n,2}(x) + 2a^4_n \xi_{n,4}(x) + o(a^6_n)] + a^4_n \xi_{n,2}^2(x) + o(a^6_n) \]  
(2.9)

\[ = a^4_n \xi_{n,2}^2(x) + O\left( \frac{a^6_n}{n} \right) \]

in view of \( \xi_{n,0}(x) = o(n^{-1}) \), where \( \xi_{n,2}(x) = \frac{1}{2} \tilde{H}_n^{(2)}(x) \mu_2(K) \).

Thus, from (2.5), (2.8) and (2.9),

\[ \text{MSE}(p_{n2}(x)) = d^2_{12}(x)n^{-1} \left[ n^{-1} \sum_{i=1}^{n} F_i(x)(1 - F_i(x)) - a_n \tilde{H}_n^{(1)}(x) \psi_1(K) \right. \]
\[ + a^2_n \tilde{H}_n^{(2)}(x) \psi_2(K) - n^{-1} \sum_{i=1}^{n} F_i(x) F_i(2)(x) \mu_2(K) \left. \right] \]
\[ + d_{12}^{-2}(x)\xi_{n,0}(x)\left[\xi_{n,0}(x) + 2a_n^2 \xi_{n,2}(x) + 2a_n^4 \xi_{n,4}(x)\right] \\
+ a_n^4 \xi_{n,2}^2(x) + O\left(\frac{a_n^4}{n}\right) \]

\[ = d_{12}^{-2}(x)[n^{-1} \sum_{i=1}^{n} F_i(x)(1 - F_i(x))/n - \frac{a_n}{n} F_n(1)\psi_1(K) + a_n^4 \xi_{n,2}^2(x)] \]

\[ + O\left(\frac{a_n^2}{n}\right) \]

as \( \xi_{n,0}(x) = \bar{H}_n(x) - H(x) = o(n^{-1}) \) as \( n \to \infty \)

\[ n^{-1} \sum_{i=1}^{n} F_i(x)(1 - F_i(x)) = n^{-1} \sum_{i=1}^{n} F_i(x) - n^{-1} \sum_{i=1}^{n} F_i^2(x) \]

\[ = \bar{H}_n(x) - \bar{H}_n^2(x) - n^{-1}(\sum_{i=1}^{n} F_i^2(x) - n\bar{H}_n^2(x)) \]

\[ = \bar{H}_n(x)(1 - \bar{H}_n(x)) - (n^{-1} \sum_{i=1}^{n} (F_i(x) - \bar{H}_n(x))^2 \]

\[ = \bar{H}_n(x)(1 - \bar{H}_n(x)) - V_{nF}(x) \] (2.10)

where \( V_{nF}(x) = n^{-1} \sum(F_i(x) - \bar{H}_n(x))^2 > 0 \) and by considering the terms containing \( \frac{a_n}{n} \)
as of higher order,

\[ \text{MSE}(p_{n,2}(x)) = d_{12}^{-2}(x)[\frac{R_n(x)[1-R_n(x)]}{n} - \frac{V_{nF}(x)}{n}] - d_{12}^{-2}(x)[\frac{a_n}{n} \bar{H}_n(1)\psi_1(K)] \]

\[ + \frac{a_n^4}{4} \bar{H}_n^2(x)\mu_{2}^2(K)] + O(\xi_{n,0}(x) a_n^2) + o(a_n^4) \]

**Corollary 2.1:** Under the conditions of Theorem 2.1 on \( \{F_i(x)\} \),

\[ \text{MSE}(p_{n,1}(x)) = d_{12}^{-2}(x)[\frac{R_n(x)[1-R_n(x)]}{n} - \frac{V_{nF}(x)}{n}] + O(n^{-2}) \]

where \( V_{nF}(x) = n^{-1} \sum(F_i(x) - \bar{H}_n(x))^2 > 0 \)

**Proof:** This proof follows exactly the similar line of argument as for the proof of Theorem 2.1.

### 2.2. Asymptotic Normality of \( p_{n,1}(x) \) and \( p_{n,2}(x) \)

We now consider the limiting distribution of BJ estimators \( p_{n,j}(x), j=1,2 \) of \( p \) using Lyapunov CLT to the sequence \( \{Z_{2i}(x)\} \) of independent random variables in the following Theorem.

**Theorem 2.2:** Under the conditions AI – AIII on \( \{F_i\} \), the kernel function \( k \), and the sequence \( \{a_n\} \) for each fixed \( x \),

\[ \sqrt{n} \left( p_{n,2}(x) - p \right) \xrightarrow{L} N(0, \frac{\tau^2}{d_{12}^2(x)}) \text{ as } n \to \infty \]

where \( \tau^2 = \lim_{n \to \infty} [\bar{F}_n(x)(1 - \bar{F}_n(x)) - V_{nF}(x)] \), \( V_{nF}(x) = n^{-1} \sum(F_i(x) - \bar{F}_n(x))^2 > 0 \)

**Proof:** Note from (2.4)

\[ d_{12}(x)[p_{n,2}(x)-p] = n^{-1} \sum_{i=1}^{n} Z_{2i}(x) + \tau_{n2}(x) \]
with \( \tau_{n2}(x) = n^{-1} \sum_{1}^{n} [E K(\frac{x-X_{i}}{a_{n}}) - H(x)] \)

\[ = \bar{H}_{n}(x) - H(x) + \frac{a_{n}^{2}}{2} \bar{H}^{(2)}_{n}(x) \mu_{2}(K) + o(a_{n}^{2}) \rightarrow 0 \text{ as } n \rightarrow \infty \]

\[ Z_{2i}(x) = K(\frac{x-X_{i}}{a_{n}}) - E K(\frac{x-X_{i}}{a_{n}}), \quad |Z_{2i}(x)| \leq 2 ||K|| = M < \infty \]

\[ \sigma_{Z_{2i}}^{2} = \text{Var} Z_{2i}(x) \]

\[ \frac{s_{n2}^{2}}{n} = \sum_{1}^{n} \sigma_{Z_{2i}}^{2} = n^{-1} \sum_{1}^{n} [F_{i}(1-F_{i}(x)) - a_{n} F_{i}^{(1)}(x) \psi_{1}(K) + O(a_{n}^{2})] \]

\[ = \bar{H}_{n}(x)(1-\bar{H}_{n}(x)) - V_{nf}(x) - a_{n} \bar{H}_{n}^{(1)}(x) \psi_{1}(K) + O(a_{n}^{2}/n) \]

\[ s_{n2}^{2} = O(n) \]

In order to apply Lyapunov CLT to the sequence \( \{Z_{2i}(x)\} \), consider the Lyapunov condition

\[ \frac{1}{s_{n2}^{3}} \sum_{1}^{n} E |Z_{i}|^{3} = \frac{n^{-1}}{s_{n2}^{3}} \left[ n^{-1} \sum_{1}^{n} E |Z_{2i}(x)|^{3} \right] \]

\[ = O\left( \frac{n^{3/2}}{n} \right) \rightarrow 0 \text{ as } n \rightarrow \infty \]

Now, Lyapunov condition is satisfied, and Lyapunov CLT to the sequence \( \{Z_{2i}(x)\} \) holds. As \( \tau_{n2}(x) \rightarrow 0 \) as \( n \rightarrow \infty \)

\[ n^{-1/2} \sum_{1}^{n} Z_{2i}(x) \xrightarrow{L} N(0,1) \]

\[ \frac{s_{n2}}{n^{1/2}} \rightarrow [H(x)(1-H(x)) - V(x)]^{1/2} = \tau \]

where \( V(x) = \lim_{n \rightarrow \infty} V_{F}(x) = \lim_{n \rightarrow \infty} n^{-1} \sum_{1}^{n} (F_{i}(x) - \bar{F}_{n}(x))^{2} \)

i.e.

\[ d_{12}(x) \sqrt{n} \left( p_{n,2}(x) - p \right) \xrightarrow{L} N(0, \tau^{2}) \]

\[ \sqrt{n} \left( p_{n,2}(x) - p \right) \xrightarrow{L} N(0, (\tau/d_{12}(x))^{2}) \]

Thus, the Theorem is proved.

**Corollary 2.2:** Under the conditions of Theorem 2.1 on \( \{F_{i}(x)\} \),

\[ \sqrt{n} \left( p_{n,1}(x) - p \right) \xrightarrow{L} N(0, \frac{\tau^{2}}{d_{12}(x)}) \text{ as } n \rightarrow \infty \]

**Proof:** This proof follows exactly the similar line of argument as for the proof of Theorem 2.2. \( \blacksquare \)

**2.3. Rates of strong convergence of \( p_{n,1}(x) \) and \( p_{n,2}(x) \)**

We now establish a.s. convergence of the BJ type estimators \( p_{n,1}(x) \) and \( p_{n,2}(x) \) defined in (1.8) under non-i.i.d. set-up in the following result:

**Theorem 2.3:** Under the conditions of Theorem 2.1,

1. \( p_{n,2}(x) - p = O\left( \frac{\log n}{n} \right)^{1/2} \text{ a.s.} \)
II. \( \hat{\lambda}_n(x) = \tilde{H}_n(x) - H(x) = O\left(\frac{\log n}{n}\right)^{1/2} \) a.s. as \( n \to \infty \)

**Proof:** Note that from (2.4) and (2.8)

\[
E(Z_{2i}(x)) = \sigma_{2i}^2 = F_i(x)(1-F_i(x)) - a_n F_i^{(1)}(x)\psi_1(K) + O(a_n^2)
\leq \frac{1}{4} + |\psi_1(K)| = C < \infty
\]

\[
\sigma_{n2}^2 = \frac{1}{n} \sum E[Z_{2i}(x)] = \frac{\sum F_i(x)(1-F_i(x))}{n} - \frac{a_n \psi_1(K) \sum F_i^{(1)}(x)}{n} + O(a_n^2)
= \tilde{H}_n(x)(1-\tilde{H}_n(x)) - V_{nF}(x) - a_n \tilde{H}_n^{(1)}(x)\psi_1(K) + O(a_n^2) < \infty
\]

By applying Bernstein (1946) inequality to \( \{Z_{2i}(x)\} \) with \( M=2 \),

\[
P(n^{-1} \sum Z_{n2}(x) > t) \leq \exp\left( -\frac{n^{-1} \sum Z_{n2}(x)}{\sigma_{n2}^2} \right)
\]

setting \( t = \left(\frac{4C \log n}{n}\right)^{1/2} \)

\[
P(n^{-1} \sum Z_{n2}(x) > t) \leq \exp\left[ -\frac{n \log n}{4C \log n} \right]
= \exp\left[ \frac{-2 \log n}{1 + \frac{4 \log n}{C}} \right]
\leq n^{-2} \quad \text{for sufficiently large.}
\]

\[
\Rightarrow \sum_{n \geq 1} P(\tilde{Z}_{n2} > t) \leq \sum_{n \geq 1} n^{-2} < \infty
\]

By Borel–Cantelli lemma, we conclude that \( \tilde{Z}_{n2} = O\left(\frac{\log n}{n}\right)^{1/2} \) as \( n \to \infty \). Therefore, as \( \tau_{n2}(x) \to 0 \) as \( n \to \infty \),

\[
(p_{n,2}(x) - p) d_1(x) = \tilde{Z}_{n2} + \tau_{n2} \xrightarrow{a.s.} O\left(\frac{\log n}{n}\right)^{1/2}
\]

i.e. \( p_{n,2}(x) - p = O\left(\frac{\log n}{n}\right)^{1/2} \) a.s. for each \( x \) as \( n \to \infty \).

(II) is an immediate consequence of part(I). Hence the result follows.

**Corollary 2.3:** Under the conditions of Theorem 2.1 on \( \{F_i(x)\} \),

I. \( p_{n,1}(x) - p = O\left(\frac{\log n}{n}\right)^{1/2} \) a.s.

II. \( \hat{\lambda}_n(x) = \tilde{H}_n(x) - H(x) = O\left(\frac{\log n}{n}\right)^{1/2} \) a.s. as \( n \to \infty \)

**Proof:** The proof follows exactly the similar line of argument as for the proof of Theorem 2.3.
3. Optimal bandwidth $a_{n,\text{opt}}$

We select the optimal $a_{n,\text{opt}}$ as that $a_n$ for which $\text{MSE} (p_{n,2}(x))$ is the minimum. Solving the equation $\frac{\partial \text{MSE} (p_{n,2}(x))}{\partial a_n} = 0$ for $a_n$:

$$\frac{\partial M}{\partial a_n} = 0 = -\frac{1}{n} \xi_{n,1}(x) + 4a_n^3 \xi_{n,2}^2(x)$$

so that

$$a_{n,\text{opt}} = \left[ \frac{\xi_{n,1}(x)}{4\xi_{n,2}^2(x)} \right]^{1/3} \cdot n^{-1/3} \quad (3.1)$$

where $\xi_{n,1}(x) = \overline{H}_n^{(1)}(x) \psi_1(K), \xi_{n,2}(x) = \frac{1}{4} \overline{H}_n^{(2)}(x) \mu_2^2(K), \psi_1(K) = 2 \int tK(t) dK(t)$

4. Comparisons between the estimators

We first compare the performance of the proposed smoothed estimator $p_{n,2}(x)$ with Boes-James type estimator $p_{n,1}(x)$, when $H_1(x), H_2(x)$ are known based on the minimum mean square error (MSE) criterion under non-i.i.d. set-up. Note that from Theorem 2.1 and Corollary 2.1,

$$\text{MSE} (p_{n,2}(x)) < \text{MSE} (p_{n,1}(x))$$

$$\text{If } \frac{a_n}{n} \overline{H}_n^{(1)}(x) \psi_1(K) > \frac{1}{4} \overline{H}_n^{(2)}(x) \mu_2^2(K)$$

$$\text{If } a_n \overline{H}_n^{(1)}(x) \psi_1(K) > na_n^4 \left[ \frac{1}{4} \overline{H}_n^{(2)}(x) \mu_2^2(K) \right] \quad (4.1)$$

for finite values of $n$. Since both terms on the left side of the above inequality are always positive and in view of $na_n^4 \rightarrow 0$ for moderate $n$, (4.1) holds. The gain in precision of $p_{n,2}(x)$ over $p_{n,1}(x)$ is defined as

$$\frac{\text{MSE} (p_{n,1}(x)) - \text{MSE} (p_{n,2}(x))}{\text{MSE} (p_{n,1}(x))} \times 100.$$ 

5. Monte Carlo Simulation

A simulation study is carried out in the estimation of $p$ by $p_{n,j}(x); j=1,2$ when two component distributions are known and are estimated by using empirical distribution function and kernel distribution function for Normal and Exponential populations. The procedure is given in appendix A.
Table 5.1. Simulation results of $p_{n,j}(x_0)$ and $p_{n,j}(x_0)$ for different sets $N$ of sample size $n$ with $p=0.5$ and $X_0 = -2, -1, -0.5, 0.5, 1, 2$ and $X_0 = 0.2, 0.3, 0.33, 0.4, 0.5, 0.6$ for Exponential population

<table>
<thead>
<tr>
<th>$p=0.5$</th>
<th>$N$</th>
<th>$H_1(x)=N(\beta t_1,0.5^2)$, $H_2(x)=N(\beta t_1,3^2)$, $H(x)=N(\beta t_1, (2.151)^2)$</th>
<th>$p=0.5$</th>
<th>$H_1(x)=\text{Exp}(2)$, $H_2(x)=\text{Exp}(3)$, $H(x)=\text{Weibull}(1.25,k=0.5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$p_{n,1}(x_0)$</td>
<td>$p_{n,2}(x_0)$</td>
<td>$MSE$</td>
</tr>
<tr>
<td>$X_0=2$</td>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.538</td>
<td>0.119</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.525</td>
<td>0.140</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.525</td>
<td>0.144</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.516</td>
<td>0.148</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.531</td>
<td>0.154</td>
<td>0.039</td>
</tr>
<tr>
<td>$X_0=1$</td>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.369</td>
<td>0.228</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.431</td>
<td>0.248</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.383</td>
<td>0.249</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.376</td>
<td>0.252</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.460</td>
<td>0.300</td>
<td>0.106</td>
</tr>
<tr>
<td>$X_0=-0.5$</td>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.463</td>
<td>0.388</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.446</td>
<td>0.360</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.438</td>
<td>0.342</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.475</td>
<td>0.335</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.461</td>
<td>0.325</td>
<td>0.068</td>
</tr>
<tr>
<td>$X_0=0.5$</td>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.58</td>
<td>0.16</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.52</td>
<td>0.16</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.51</td>
<td>0.19</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.47</td>
<td>0.17</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.47</td>
<td>0.18</td>
<td>0.056</td>
</tr>
<tr>
<td>$X_0=1$</td>
<td>n=12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.439</td>
<td>0.127</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>0.512</td>
<td>0.135</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.525</td>
<td>0.169</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>0.494</td>
<td>0.163</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.518</td>
<td>0.153</td>
<td>0.047</td>
</tr>
</tbody>
</table>

Y. S. Ramakrishnaiah, et al.: On the smoothed parametric...
6. Comments

In the paper it is shown that when the component normal populations with parameters are \( N(\beta_{t_i}, 0.25) \), \( N(\beta_{t_i}, 9) \) and the mean value of estimate \( p_{n,1}(x_0) \) and \( p_{n,2}(x_0) \) is close to its actual value \( p \). The simulation results show that \( \hat{MSE} \) for smoothed parametric estimator is less than that of unsmoothed estimator for different values of \( x \), uniformly for all samples. Thus, the smoothed estimator is a better estimator in terms of minimum MSE when compared to BJ estimator. The average gain in efficiency due to smoothing lies between 19% to 88% for different sets \( N \) of size \( n \).

REFERENCES


APPENDIX A

Random samples of sizes $n_1=6$ and $n_2=6$ are generated from the two component mixtures of the normal populations with parameters \((\mu_1, \mu_2) = (\beta t_1, \beta t_2)\) and \((\sigma_1^2, \sigma_2^2) = (0.5^2, 3^2)\) and with parameters \((\theta_1, \theta_2) = (2, 3)\) for Exponential populations. The mixed sample of size $n = n_1 + n_2 = 12$ is generated from the normal population with parameters \((\mu = p\mu_1 + q\mu_2, \sigma^2 = p\sigma_1^2 + q\sigma_2^2)\), and in the case of Exponential population, the mixed sample is drawn from Weibull population with shape parameter $k$ less than 1. Since Weibull distribution with shape parameter $k<1$ arises as a mixture of Exponential distributions (Jewel 1982), the samples of sizes $n, n_1, n_2$ are independent. Taking $p=q=0.5, \beta=0.1$ and $t_i = \mp \frac{1}{n_0^p}; j = 1, 2, \delta =1.5$ are selected in such a way that $\sum t_i = 0$ and $\sum t_i^2 \to 0$. The present simulation study is to estimate parametric estimators such as $p_{n,1}(x)$ and $p_{n,2}(x)$ for $x=x_0$ defined as follows.

\[
p_{n,1}(x_0) = \frac{\hat{H}_n(x_0) - H_2(x_0)}{H_1(x_0) - H_2(x_0)} \quad \text{and} \quad p_{n,2}(x_0) = \frac{\hat{H}_n(x_0) - H_2(x_0)}{H_1(x_0) - H_2(x_0)}
\]

where $\hat{H}_n(x_0), \hat{H}_n(x_0)$ are estimated by the usual empirical and kernel-based distribution functions and $H_j(x), j=1,2$ such as

\[
\hat{H}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x), \text{ if } I(X_i \leq x), \text{ assign } 1 \text{ otherwise } 0.
\]

\[
\hat{H}_n(x_0) = n^{-1} \sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{a_n}\right); H_j(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} F_{ij}(x_0).
\]

Here, we used the Epanechnikov kernel function as $k(u) = \frac{3}{4} (1 - u^2); |u| \leq 1$ for Normal distribution, and for Exponential distribution we used Gaussian kernel function as $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$. The distribution function of the Epanechnikov kernel function is

\[
K\left(\frac{x_0 - X_i}{a_n}\right) = \frac{3}{4} \left[\frac{x_0 - X_i}{a_n} - \frac{1}{3} \left(\frac{x_0 - X_i}{a_n}\right)^3 + \frac{2}{3}\right]
\]

Thus, the estimators becomes

\[
p_{n,1}(x_0) = \frac{\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x_0) - \frac{1}{n_2} \sum_{i=1}^{n_2} F_{2i}(x_0)}{\frac{1}{n_1} \sum_{i=1}^{n_1} F_{1i}(x_0) - \frac{1}{n_2} \sum_{i=1}^{n_2} F_{2i}(x_0)}
\]

\[
p_{n,2}(x_0) = \frac{n^{-1} \sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{a_n}\right) - \frac{1}{n_2} \sum_{i=1}^{n_2} F_{2i}(x_0)}{\frac{1}{n_1} \sum_{i=1}^{n_1} F_{1i}(x_0) - \frac{1}{n_2} \sum_{i=1}^{n_2} F_{2i}(x_0)}
\]

(5.1)

where $F_{ij}(x)$ are the cumulative distribution functions of Normal distribution and Exponential distribution.
All computations are done by using MS Excel, and the procedure is as follows.
1. Generate \( n \) uniform random numbers between \( (0,1) \).
2. Generate the cumulative distribution function of Normal and Exponential distribution by taking different means \( \beta_{tji} \) and variances \( \sigma_j^2; \ j=1,2 \) at \( x = x_0 \).
3. Generate the mixed normal and Weibull observations by taking different means \( \beta_i = p\beta_{1i} + q\beta_{2i} \) and variance \( \sigma^2 = p\sigma_1^2 + q\sigma_2^2 \) and \( \theta = p\theta_1 + q\theta_2 \) respectively with \( k<1 \).
4. Calculate (5.1) by taking \( X_0 = -2, -1, -0.5, 0.5, 1, 2 \) related to Normal distribution and \( X_0 = 0.2, 0.3, 0.33, 0.4, 0.5, 0.6 \) related to Exponential distribution.

Generate \( N=100 \) mixed and component independent sample sets with sample sizes \( n=12 \) and \( n_1=6 \) and \( n_2=6 \), so that \( n = n_1+n_2 \) and calculate \( p_{n,1}(x_0) \) and \( p_{n,2}(x_0) \), their mean values \( \bar{p}_{n,j}(x_0) = \frac{1}{N} \sum_{i=1}^{N} p_{n,j}(x_0) \) and their mean square errors \( MSE \ (p_{n,j}(x_0)) = \frac{1}{N} \sum_{i=1}^{N} (p_{n,j}(x_0) - \bar{p}_{n,j}(x_0))^2 ; \ j=1,2 \). Sets are ignored when \( p \geq 1 \).

The results are presented in table 5.1.