Journal of Sciences, Islamic Republic of Iran 24(1): 55-62 (2013) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Some Asymptotic Results of Kernel Density Estimator in Length-Biased Sampling

M. Ajami,[1] V. Fakoor,[1,*] and S. Jomhoori[2]

[1]*Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Islamic Republic of Iran*
[2]*Department of Statistics, Faculty of Sciences, University of Birjand, Birjand, Islamic Republic of Iran*

Received: 10 September 2012 / Revised: 6 February 2013 / Accepted: 6 May 2013

Abstract

In this paper, we prove the strong uniform consistency and asymptotic normality of the kernel density estimator proposed by Jones [12] for length-biased data.The approach is based on the invariance principle for the empirical processes proved by Horváth [10]. All simulations are drawn for different cases to demonstrate both, consistency and asymptotic normality and the method is illustrated by real automobile brake pads data.

Keywords: Asymptotic normality; Length-biased; Strong consistency; Strong Gaussian approximation

Introduction

Given a distribution function (d.f.) F, we say a random variable (r.v.) Y has the length-biased distribution of F if the d.f. of Y is given by

$$G(t) = \frac{1}{\mu}\int_0^t x dF(x), \quad t \geq 0, \tag{1.1}$$

where $\mu = \int_0^\infty x dF(x)$, and μ is assumed to be finite. In the case that F has a density f, with respect to the Lebesgue measure, (1.1) can be written as

$$G(t) = \frac{1}{\mu}\int_0^t x f(x) dx,$$

and hence the density of Y is given by

$$g(t) = \frac{tf(t)}{\mu}, \quad t \geq 0.$$

The phenomenon of length-bias was first tackled in the context of anatomy by Wicksell [26] as what he called the corpuscle problem. Length-biased sampling was later systematically studied by McFadden [15], Blumenthal [3], then by Cox [6], in the context of estimation of the distribution of fiber lengths in a fabric.

Length-biased data arise in many practical situations, including econometrics, survival analysis, renewal processes, biomedicine and physics. For instance, if X represents the length of an item and the probability of this item selected in the sample is proportional to its length, then the distribution of the observed length is length-biased. In cross-sectional studies in survival analysis, e.g. often the probability of being selected, for a particular subject, is proportional to his/her survival time. Interesting applications of length-biased data can be found in Cox [6], Patil and Rao [18, 19],Chen, et. al [5], Colman [4], Huang and Qin [11] and Vardi [24].

The distribution function,G, is from a slightly different perspective, the distribution of the randomly

* Corresponding author, Tel.: +98(511)8828605, Fax: +98(511)8828605, E-mail: fakoor@math.um.ac.ir

left truncated r.v.'s Y , in the stationary assumption. If the incidence rate of the event has not changed over time, a stationary assumption might reasonably describe the incidence of the event.This is equivalent to assume that the randomly left truncation induced by the sampling is uniform (Wang, [25]).

Throughout this paper, we assume that G is continuous on $\mathsf{R}^+ = [0,\infty)$, from which it follows that F is also continuous. An elementary calculation shows that F is determined uniquely by G, namely

$$F(t) = \mu\int_0^t y^{-1}dG(y), \quad t \geq 0.$$

Let $Y_1,\ldots,Y_n$ be a sample from G. The empirical estimator of *F* can be written in the form of

$$F_n(t) = \mu_n \int_0^t y^{-1}dG_n(y), \tag{1.2}$$

where

$$\mu_n^{-1} = \int_0^\infty y^{-1}dG_n(y).$$

G_n is an empirical estimator of G given by

$$G_n(t) = \frac{1}{n}\sum_{i=1}^n I(Y_i \leq t),$$

where $I(A)$ denotes the indicator of the event *A*.

The kernel estimator of a real univariate density function *f* introduced by Rosenblatt [20] is

$$f_n(t) = \sum_{i=1}^n \frac{1}{nh_n} K\left(\frac{t-X_i}{h_n}\right),$$

where $X_1,\ldots,X_n$ are independent observations from f, K is a kernel function and h_n is a sequence of (positive) "bandwidths" tending to zero as $n\to\infty$. Parzen [17] showed that under some mild smoothness conditions on K (and f), $f_n(t)$ is in any respect a consistent estimator of $f(t)$ for each $t \in \mathsf{R}$. The weak and strong uniform consistency properties of f_n have been considered by several authors, including Nadaraya [16], Schuster [21] and Van Ryzin [23]. In these papers the condition placed on the bandwidth for the strong uniform consistency includes $\sum \exp(-cnh_n^{\,2}) < \infty$ for all positive c. Silverman [22] established the strong uniform consistency for $f_n - f$ using the strong approximation technique developed by Komlós, et al. [13] for the ordinary empirical process. Bhattacharyya, et al. [2] proposed the following estimator for length-biased data $Y_1,\ldots,Y_n$

$$f_n^*(t) = \frac{\mu_n}{t}\sum_{i=1}^n \frac{1}{nh_n} K\left(\frac{t-Y_i}{h_n}\right). \tag{1.3}$$

Where $K(\cdot)$ is asymmetric kernel function and $\{h_n, n\geq 1\}$ is a sequence of positive bandwidths satisfying $h_n \to 0$ and $nh_n \to \infty$ as $n\to\infty$, which controls the degree of smoothness of the estimator. They proved consistency and asymptotic normality for f_n^*. Jones [12] proposed the following estimator

$$f_n(t) = \frac{\mu_n}{nh_n}\sum_{i=1}^n \frac{1}{Y_i} K\left(\frac{t-Y_i}{h_n}\right), \tag{1.4}$$

which has various advantages with respect to (1.3). It is a probability density function, it is particularly better behaved near zero, it has better asymptotic mean integrated squred error properties and it is more readily extendable to related problems such as density derivative estimation.

An interesting overview of nonparametric contributions to the literature on estimating problems when the observations are taken from weighted distributions can be found in Cristóbal and Alcalá [7]. The asymptotic results on sharp minimax density estimation for length-biased data were derived by Efromovich [8]. By using rejection sampling techniques, Guillamónet, *et al*. [9] gave an alternative estimator for the density function *f*. Ajami, *et al*. [1] also estimated the bandwidth parameter according to a Bayesian approach. They proved the strong consistency of the estimator (1.4) by applying Bayesian length-biased and compared Bayesian and the cross validated least square for estimate of the length-biased together via some simulation studies.

By equation (1.2), it is easy to see that

$$f_n(t) = \frac{1}{h_n}\int K\left(\frac{t-u}{h_n}\right)dF_n(u). \tag{1.5}$$

This paper tries to study the strong uniform consistency as well asasymptotic normality of the kernel density estimator proposed by Jones [12]. Our approach,first of all, is applying a strong approximation technique to establish the strong uniform consistency and asymptotic normality of $f_n - \tilde{f}_n$, where

$$\tilde{f}_n(t) = \frac{1}{h_n}\int_0^\infty K\left(\frac{t-s}{h_n}\right)dF(s).$$

Now, for the sake of simplicity, the assumptions used in this paper are as follows.

Assumptions

1. The kernel function K is symmetric, of bounded variation on $(-1,1)$. In addition, $K(t)=0$ if $t \notin (-1,1)$ and satisfies the following conditions:

$$\int_{-1}^{1} K(t)dt = 1,$$

$$\int_{-1}^{1} tK(t)dt = 0,$$

$$\int_{-1}^{1} t^2 K(t)dt = m \neq 0,$$

$$\int_{-1}^{1} |dK(t)| = V_K < \infty.$$

2. $\tau := \inf[t : F(t) < 1] < \infty$

3. $\int_0^{\infty} u^{-2} G^{1/r}(u)du < \infty$, for some $r > 2$.

4. $f(t) = O(t)$ as $t \to 0$.

This paper is organized as follows. In the next section, we recall some important and useful results in the length-biased model and prove a necessary modulus of continuity of the Gaussian process for proving the main results. In the main results section, we provide some asymptotic behaviors of the suggested estimator of kernel density estimator (1.5). Some simulations are drawn to grant further support of our theoretical results regarding to the consistency as well as the asymptotic normality. Besides there are some results illustrated by the real data in the application section. Finally, the proofs of the main results are postponed to in the proofs section, where some auxiliary results are also proved.

Preliminaries

Strong approximation for $\alpha_n(t) = \sqrt{n}[F_n(t) - F(t)]$ will be derived from the well-known approximations of the empirical process

$$\beta_n(t) = \sqrt{n}[G_n(t) - G(t)], \quad t \geq 0.$$

Without loss of generality, we can assume that our probability space (Ω, A, P) is so rich that the approximation

$$\sup_{t \geq 0} |\beta_n(t) - k(t,n)| = O\left(n^{-\frac{1}{2}}(\log n)^2\right) \quad a.s.,$$

of Komlós, *et al.* [13] holds, where $k(t,n)$ is a two parameter Gaussian process with zero mean and covariance function

$$E[k(x,n)k(y,m)] = (mn)^{-1/2}(m \wedge n)[G(x \wedge y) - G(x)G(y)] \quad (a \wedge b = \min(a,b)).$$

Using $k(t,n)$, Horváth [10] defined the process $B(t,n)$ to approximate α_n, such that

$$B(t,n) = \mu \int_0^t y^{-1} dk(y,n) - \mu F(t) \int_0^{\infty} y^{-1} dk(y,n).$$

It is easy to check that $\{B(t,n), 0 \leq t < \infty, n \geq 1\}$ is a Gaussian process with zero mean and covariance function

$$\begin{aligned} &E[B(x,n)B(y,m)] \\ &= (mn)^{-1/2}(m \wedge n)[\sigma(x \wedge y) \\ &- F(x)\sigma(y) - F(y)\sigma(x) + F(x)F(y)\sigma], \end{aligned} \tag{2.1}$$

where

$$\sigma(t) = \mu^2 \int_0^t y^{-2} dG(y),$$

and

$$\sigma = \lim_{t \to \infty} \sigma(t) = \mu^2 \int_0^{\infty} y^{-2} dG(y).$$

Let $\{W(u,v), \ u,v \geq 0\}$ denoting a two-parameter Wiener process. By (2.1) the following representation holds

$$\begin{aligned} \{n^{1/2}B(t,n), \ t \geq 0, n \geq 1\} &\overset{D}{=} \{W(\sigma(t), n) \\ &- F(t)W(\sigma, n), \ t \geq 0, n \geq 1\}, \end{aligned} \tag{2.2}$$

where $\overset{D}{=}$ denotes an equal in distribution.

Theorem 1. *(Horváth, [10])* S*uppose* the Assumption (3) is satisfied. On a suitably enlarged probability space, there exists a two-parameter *m*ean zero G*aussian* process $\{B(t,n), 0 \leq t < \infty, n \geq 1\}$ with covariance (2.1), such that

$$\sup_{0 \leq t < \infty} |\alpha_n(t) - B(t,n)| = O(n^{-\lambda}) \quad a.s.,$$

for any $0 < \lambda < 1/2 - 1/r$, and some $r > 2$. □

To study the strong consistency of f_n, we have also need to study the modulus of continuity of the approximating process $B(u,n)$. In the following

lemma, we prove the global modulus of continuity of the Gaussian process $B(u,n)$.

Lemma 1. *Let* $\{h_n\}$ *be a sequence of positive bandwidth for which* $h_n \to 0$ *as* $n \to \infty$. *Suppose the Assumption* (4) *holds. Then, we have*

$$\sup_{0\le t\le\tau}\sup_{-1\le u\le 1}|B(t-uh_n,n)-B(t,n)|=O\left(\sqrt{h_n\log h_n^{-1}}\right)\ a.s.$$

Proof. Since

$$\left\{\frac{W(x,n)}{\sqrt{n}};0\le x<\infty,n\ge 1\right\}\overset{D}{=} \left\{W_n(x);0\le x<\infty,n\ge 1\right\}, \tag{2.3}$$

where $W_n(x)$ is a sequence of standard Wiener processes. According to (2.2), it is enough to show that,

$$\sup_{0\le t\le\tau}\sup_{-1\le u\le 1}|W_n(\sigma(t-uh_n))-W_n(\sigma(t))| = O\left(\sqrt{h_n\log h_n^{-1}}\right)\ a.s., \tag{2.4}$$

and

$$\sup_{0\le t\le\tau}\sup_{-1\le u\le 1}|F(t-uh_n)-F(t)|\left|\frac{W(\sigma,n)}{\sqrt{n}}\right| = O(h_n)\ \ a.s. \tag{2.5}$$

Let $t\in[0,\tau]$ and $M_f=\sup_{0\le t\le\tau}f(t)$. By the boundedness of f on $[0,\tau]$, it follows for the Mean Value Theorem that $|F(t-uh_n)-F(t)|\le M_f h_n$. So, we have (2.5).

To prove (2.4), first note that

$$\sigma(t-uh_n)-\sigma(t)=\mu^2\int_{t\wedge(t-uh_n)}^{t\vee(t-uh_n)}y^{-2}dG(y)$$

$$=\mu\int_{t\wedge(t-uh_n)}^{t\vee(t-uh_n)}y^{-1}dF(y).$$

By Assumption (4), for $u\in[-1,1]$

$$|\sigma(t-uh_n)-\sigma(t)|\le Ch_n,$$

where C is a positive constant. Moreover,

$$\sup_{0\le t\le\tau}\sup_{-1\le u\le 1}|W_n(\sigma(t-uh_n))-W_n(\sigma(t))|\le \sup_{0\le x\le\sigma}\sup_{0\le y\le Ch_n}|W_n(x+y)-W_n(x)|=O\left(\sqrt{h_n\log h_n^{-1}}\right)\ a.s.,$$

where the last equality is proved along the line of the equation (2.4) of Zhang [27]. So, we obtain (2.4). Now, by (2.3), (2.4) and (2.5), we get the result. □

Results

1- Strong Uniform Consistency

In the following theorem, we prove strong uniform consistency of f_n.

Theorem 2. *Let* h_n *be a sequence of positive bandwidths tending to zero as* $n\to\infty$. *Assumptions* (1) - (4) *hold and*

$$\frac{\log n}{n^{\frac{1}{2}+\lambda}h_n}\to 0 \quad as \quad n\to\infty, \tag{3.1}$$

for any $0<\lambda<1/2-1/r$, and some $r>2$. Then

$$\lim_{n\to\infty}\sup_{0\le t\le\tau}|f_n(t)-f(t)|=0 \quad a.s.$$

Proof. See Section 4. □

Remark 1. *If the bandwidth* h_n *is chosen to be* $h_n\sim\alpha n^{-\beta}$ *with* $\alpha>0$ *and* $0<\beta<\frac{1}{2}+\lambda$, *then condition (3.1) is satisfied.*

2- Asymptotic Normality

In the following theorem, we study the asymptotic normality of f_n.

Theorem 3. *Suppose* f *is continuous at* $t\in(0,\tau]$ *and Assumptions (1)-(3) are fulfilled. Let* $h_n\to 0$ *and* $n^{2\lambda}h_n\to\infty$ *as* $n\to\infty$, *for any* $0<\lambda<\frac{1}{2}-\frac{1}{r}$, *for some* $r>2$.

Then for $t\in(0,\tau]$

$$\sqrt{nh_n}[f_n(t)-\tilde{f}_n(t)]\overset{D}{\to}N(0,\gamma^2(t)) \quad as\ n\to\infty,$$

where

$$\gamma^2(t)=\frac{\mu f(t)}{t}\int_{-1}^{1}K^2(u)du. \tag{3.2}$$

Proof. See Section 4. □

Corollary 1. *In addition to the conditions in Theorem 3, if f has a bounded derivative in a neighborhood of t and* $nh_n^3 \to 0$ *as* $n \to \infty$, *then*

$$\sqrt{nh_n}[f_n(t) - f(t)] \xrightarrow{D} N(0, \gamma^2(t)) \tag{3.3}$$

Proof. Suppose $|f'(s)| \le M_t$ for any s in a neighborhood of t, where M_t is a constant depending only on t Applying the Mean Value Theorem with $\xi_n \in (\min(t, t-h_n u), \max(t, t-h_n u))$ gives

$$\begin{aligned}
&\sqrt{nh_n}\left|f_n(t) - f(t)\right| \\
&= \sqrt{nh_n}\left|\int_{-1}^{1} K(u)[f(t-h_n u) - f(t)]du\right| \\
&= \sqrt{nh_n}\left|\int_{-1}^{1} K(u)[f'(\xi_n)(-h_n u)]du\right| \\
&\le M_t \sqrt{n} h_n^{3/2} \int_{-1}^{1} |uK(u)|du \to 0,
\end{aligned} \tag{3.4}$$

as $n \to \infty$. Combining (3.4) and Theorem 3 completes the proof. □

Corollary 2. Using Corollary 1, it is possible to construct confidence interval for $f(t)$. A plug-in estimate

$$\gamma_n^2(t) := \frac{\mu_n f_n(t)}{t} \int_{-1}^{1} K^2(u)du$$

of the asymptotic variance $\gamma^2(t)$ can be easily obtained using (1.4). This estimator is consistent and yields a confidence interval of asymptotic level $1-\alpha$ for $f(t)$ namely,

$$[f_n(t) - \frac{\gamma_n(t)}{\sqrt{nh_n}} z_{1-\frac{\alpha}{2}}, f_n(t) + \frac{\gamma_n(t)}{\sqrt{nh_n}} z_{1-\frac{\alpha}{2}}], \tag{3.5}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the $(1-\frac{\alpha}{2})$-quantile of a standard normal distribution.

Corollary 3. In addition to the conditions in Theorem 3, if f satisfies

$$|f(t+h) - f(t)| \le C_t |h|^{\alpha}$$

for any h in a neighborhood of 0, where $\alpha \in [0,1]$ and C_t depends only on t and if $nh_n^{1+2\alpha} \to 0$ as $n \to \infty$, then we get

$$\sqrt{nh_n}[f_n(t) - f(t)] \xrightarrow{D} N(0, \gamma^2(t)),$$

where $\gamma^2(t)$ is given by (3.2).

Proof. As $n \to \infty$,

$$\begin{aligned}
&\sqrt{nh_n}\,|f_n(t) - f(t)| \\
&= \sqrt{nh_n}\left|\int_{-1}^{1} K(u)[f(t-h_n u) - f(t)]du\right| \\
&\le C_t \sqrt{nh_n^{1+2\alpha}} \int_{-1}^{1} |uK(u)|du \to 0.
\end{aligned}$$

It is completed The proof. □

Remark 2. The following corollary is the same as Proposition 5 of Guillamón, *et al.* [9]. They used assumption $E(Y^{-2}) < \infty$ which is slightly weaker than Assumption (3).

Corollary 4. In addition to the conditions in Theorem 3, if $\lambda > \frac{1}{10}$ and f is twice continuously differentiable in a neighborhood of t and the bandwidth h_n satisfies $h_n = O(n^{-1/5})$ as $n \to \infty$, then, we have

$$\begin{aligned}
&\sqrt{nh_n}(f_n(t) - f(t) \\
&-\frac{1}{2}h_n^2 f''(t)m) \xrightarrow{D} N(0, \gamma^2(t)) \quad n \to \infty,
\end{aligned}$$

where m is given in Assumptions (1).

Proof. Applying a two-term Taylor expansion gives, as $n \to \infty$,

$$\begin{aligned}
&\sqrt{nh_n}\left|f_n(t) - f(t) - \frac{1}{2}h_n^2 f''(t)m\right| \\
&= \frac{1}{2}\sqrt{n}h_n^{5/2}\left|\int_{-1}^{1} u^2 K(u) f''(\xi_n)du - f''(t)m\right| \\
&= \frac{1}{2}\sqrt{n}h_n^{5/2}\left|\int_{-1}^{1} u^2 K(u)[f''(\xi_n)du - f''(t)]du\right| \\
&\le \frac{1}{2}\sqrt{n}h_n^{5/2}\int_{-1}^{1} u^2 |K(u)||f''(\xi_n)du - f''(t)|du \\
&\to 0,
\end{aligned} \tag{3.6}$$

where $\xi_n \in (\min(t, t-h_n u), \max(t, t-h_n u))$. The proof is complete by combining (3.6) and Theorem 3.

Application

This section has two parts: The first part shows the behavior of the Jones' estimator for strong consistency and asymptotic normality and the second one deals with kernel density estimator and confidence bound for the real data.

1- Simulation

Having illustrated the behavior of the proposed method, here we present the results of a preliminary small-sample simulation study, especially those of Monte Carlo method.

In order to check the consistency of the Jones' estimator (1.4), first the graphs of unbiased density function *f and* the Jones' estimator are demonstrated in the same figure. Besides, it is supposed that the data are emanated from an unbiased model with underlying Gamma density function $f(t) = t\, e^{-t}$, $t > 0$, thus the length-biased density is $g(t) = t^2 e^{-t}/2$, $t > 0$. The estimator of μ is taken to be $\hat{\mu}_n = n(\sum_{i=1}^{n} Y_i^{-1})^{-1}$. Obviously, the density $f(t)$ satisfies the assumption (4) .For this simulation study, a sample of size 100 is taken. The Epanechnikov kernel function

$$K(t) = \frac{3}{4}(1-t^2) I_{(-1,1)}(t), \quad t > 0, \tag{4.1}$$

is used to construct a density estimator. It should be noticed that applied the Epanechnikov kernel (4.1) satisfies the assumption (1). Bandwidth parameter is chosen based on the least square cross validation method discussed in [1]. According to the Figure 1, the Jones' estimator can estimate the density function $f(\cdot)$ properly.

Now we consider the asymptotic normality property. Following this purpose, corresponding histogram and Q-Q-normal plots are illustrated. Besides by applying simulation and nonparametric Kolmogorov-Smirnov test we show that $\frac{\sqrt{nh_n}[f_n(t) - f(t)]}{\gamma(t)}$ at $t = 2$ has asymptotically standard Normal distribution, where $\gamma^2(t) = \frac{\mu f(t)}{t}\int_{-1}^{1} K^2(u)du$. Farther more, assume the data emanating from the length-biased model, with underlying Gamma density function $g(t) = t^2 e^{-t}/2$, $t > 0$, and sample size n=100. Bandwidth parameter is $h_n = n^{-\frac{2}{5}}$, which satisfies the conditions of Theorem 3 and Corollary 1. The kernel (4.1) is also used. According to the histograms and Q-Q-normal plots, we trivially notice that $\frac{\sqrt{nh_n}[f_n(2) - f(2)]}{\gamma(2)}$ has asymptotically standard Normal disribution. Furthermore Kolmogorov-Smirnov test gives the respective p-values (0.318) which suggests not to reject the Normality distribution.

2- Real Data

According to the length-biased lifetime data of 98 automobile brake pads (in 1000-km units) [13], the unbiased density function $f(.)$ is estimated. By employing (3.3), construct a 95% asymptotic confidence band for the true density defined as (3.5). The proposed estimator (1.4) can be applied to this data set with n=98. The applied bandwidth is achieved by a subjective selection method and the Epanechnikov kernel function defined in (4.1) is employed. Figure 3 reports the estimate and associated 95% confidence bands for the true density (3.5) and it proposes that the density of unbiased population may be generated from a Gamma distribution function.

Proofs

In order to make the proof easier, we need some auxiliary results and notations. The first result gives a uniform consistency of $f_n - \tilde{f}_n$. □

Lemma 2. Assuming the same conditions as in Theorem 2, we have

$$\limsup_{n\to\infty} \sup_{0\le t\le \tau} |f_n(t) - \tilde{f}_n(t)| = 0 \quad a.s.$$

Proof. By equation (1.5) and according to Theorem 1, there exists a Gaussian process $B(t,n)$ such that, for large *n* and $t \in [0,\infty)$, we have

$$f_n(t) - \tilde{f}_n(t) = -\frac{1}{\sqrt{n}h_n}\int_0^{\infty} \alpha_n(x)dK\left(\frac{t-x}{h_n}\right)$$

$$= \frac{1}{\sqrt{n}h_n}\int_{-1}^{1} B(t-uh_n, n)dK(u) + O\left(\frac{n^{-\lambda}}{\sqrt{n}h_n}\right) a.s.$$

$$= \frac{1}{\sqrt{n}h_n}\int_{-1}^{1} [B(t-uh_n, n) - B(t,n)]dK(u) + O\left(\frac{n^{-\lambda}}{\sqrt{n}h_n}\right) a.s.$$

Now Lemma 1, and (3.1) complete the proof. □

Proof of Theorem 2. Since f is continuous on $[0,\tau]$, f is uniformly continuous on $[0,\tau]$, hence by the dominated convergence theorem, it can be shown that

$$\limsup_{n\to\infty} \sup_{0\le t\le\tau} |\tilde{f}_n(t) - f(t)| = 0.$$

Therefore, Theorem 2 is a straightforward consequence of Lemma 2 and the equality

$$f_n - f = f_n - \tilde{f}_n + \tilde{f}_n - f. \qquad \square$$

To prove the Theorem 3, first we introduce some further notations. Let $\{W(t), t \ge 0\}$ be a standard Wiener process and

$$A_n(t) = \frac{1}{\sqrt{nh_n}} \int_{-1}^{1} W(\sigma(t - uh_n), n) dK(u)$$

$$B_n(t) = -\frac{1}{\sqrt{nh_n}} \int_{-1}^{1} [F(t - uh_n) - F(t)] W(\sigma, n) dK(u)$$

$$C_n(t) = \frac{1}{\sqrt{h_n}} \int_{0}^{\infty} K\left(\frac{t-u}{h_n}\right) \sqrt{\sigma'(u)} dW(u)$$

Proof of Theorem 3. By equation (1.2) and according to Theorem 1, there exists a Gaussian process $B(t,n)$ such that, for large *n* and $t \in (0,\tau]$, we have

$$\begin{aligned} &\sqrt{nh_n}[f_n(t) - \tilde{f}_n(t)] \\ &= -\frac{1}{\sqrt{h_n}} \int_0^{\infty} \sqrt{n}[F_n(x) - F(x)] dK\left(\frac{t-x}{h_n}\right) \\ &= \frac{1}{\sqrt{h_n}} \int_{-1}^{1} B(t - uh_n) dK(u) + O\left(\frac{n^{-\lambda}}{\sqrt{h_n}}\right) \end{aligned} \tag{4.1}$$

By (2.2), for each *n*,

$$\frac{1}{\sqrt{h_n}} \int_{-1}^{1} B(t - uh_n, n) dK(u) \overset{D}{=} A_n(t) + B_n(t).$$

To prove the theorem, it is enough to show that $B_n(t) = O_p(h_n^{\frac{1}{2}})$ and $A_n(t) \xrightarrow{D} N(0, \gamma^2(t))$, then applying of Slutsky's Theorem completes the proof. Since f is continuous at *t*, f is bounded in a neighborhood of *t*, i.e., there exists a constant M_t such that $f(\xi_n(u)) \le M_t$ uniformly for *u* in $(-1,1)$. Hence,

$$|B_n(t)| \le \sqrt{h_n} M_t V_K |W(\sigma, n)| / \sqrt{n}$$

and

$$E|B_n(t)| \le \sqrt{h_n} M_t V_K \sqrt{\frac{2\sigma}{\pi}}, \tag{4.3}$$

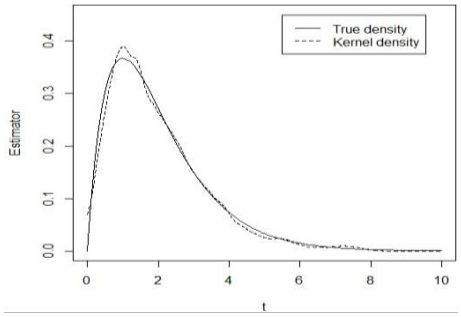

Figure 1. Unbiased density function (solid line) and Jones' estimator (dashed line).

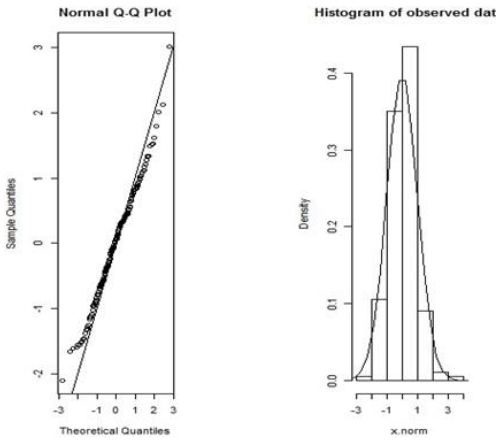

Figure 2. Histogram and Normal Q-Q plots for simulation data from Length-biased sampling of gamma distribution.

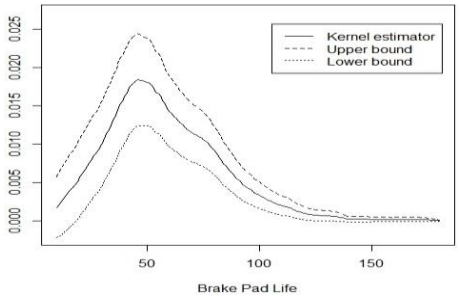

Figure 3. Kernel density estimator and confidence bounds for the lifetime of automobile brake pads.

which implies that for each fixed $t \geq 0$

$$B_n(t) = O_p(h_n^{\frac{1}{2}}).$$

Now, let

$$\theta_n^2(t) = Var(C_n(t)) = \int_{-1}^{1} K^2(u)\sigma'(t - h_n u)du.$$

By the continuity of *f* at *t*it can be shown that

$$\lim_{n\to\infty} Var(C_n(t)) = \gamma^2(t). \quad (4.4)$$

Since $C_n(t)$ is a normal random variable with mean 0 and variance $\theta_n^2(t)$, it follows from (4.4) and Slutsky's Theorem that

$$C_n(t) \xrightarrow{D} N(0, \gamma^2(t)). \quad (4.5)$$

Furthermore, since $A_n(t) \stackrel{D}{=} C_n(t)$ for each *n*, (4.5) implies that

$$A_n(t) \xrightarrow{D} N(0, \gamma^2(t)). \quad (4.6)$$

Combining (4.1), (4.2), (4.3), and (4.6) completes the proof of Theorem 3. □

Acknowledgements

The authors would like to sincerely thank anonymous referees for the careful reading of the manuscript.

References

1. Ajami, M., Fakoor, V. and Jomhoori, S. Bayesian and cross validation estimation bandwidth of kernel density function estimator for length-biased data. *Journal of Statistical Sciences.* **5(1)**: 41-60 (2011).
2. Bhattacharyya, B. B., Franklin, L. A. and Richardson, G. D. A comparison of nonparametric unweighted and length-biased density estimation of fibres. Comm. Statist. **A 17,** 3629-3644 (1988).
3. Blumenthal, S. Proportional sampling in life length studies. Technometrics. **9**: 205-218 (1967).
4. Colman, R. An introduction to mathematical sterology. Memoirs No. 3, Dept. of Theoretical Statistics, Univ. of Aarhus, Denmark (1979).
5. Chan, K.C., Chen, Y. Q. and Di, C. Z. Proportional mean residual life model for right-censored length-biased data. Biometrika. **10:** 1-6 (2012).
6. Cox, D. R. Some sampling problems in technology. New developments in survey sampling, (eds: N.L. Johnson and H. Smith). New York: Wiley (1969).
7. Cristóbal, J. A., Alcalá, J. T. An overview of nonparametric contributions to the problem of functional estimation from biased data. Test. **10**, No. 2:309-332 (2001).
8. Efromovich, S. Density estimation for biased data. Ann. Stat. **32**, No. 3: 1137-1161 (2004).
9. Guillamón, A., Navarro, J. and Ruiz, J. M. Kernel density estimation using weigthed data. Commun. Statist.- Theory Meth. **27(9)**: 2123-2135 (1998).
10. Horváth, L. Estimation from a length-biased distribution. Stat. Decis. **3:** 91-113 (1985).
11. Huang, CH. Y., Qin, J. Nonparametric estimation for length-biased and right-censored data. *Biometrika.* **98 (1):** 177-186 (2011)
12. Jones, M. C. Kernel density estimation for length biased data. Biometrika.**78**: 511-519 (1991).
13. Komlós, J., Major, P. and Tusnády, G. An approximation of partial sums of independent r.v.'s and the sample d.f. I.Z. Wahrscheinlichkeitstheorie verw.Gebiete. **32**: 111-131 (1975).
14. Lawless J.F. Statistical Models and Methods for Lifetime Data, 2nd Ed., John Wiley & Sons, New York, pp 69-70 (2003)
15. McFadden, J. A. On the lengths of intervals in a stationary point process. Journal of the Royal Society. Series B, **24**: 364-382 (1962).
16. Nadaraya, E. A. On non-parametric estimates of density functions and regression curves. Theor. Prob. Appl. **10**: 186-190 (1965).
17. Parzen, E. On estimation of a probability density function and mode. Ann. Math. Statist. **33**: 1065-1076 (1962).
18. Patil, G. P. and Rao, C. R. The weighted distributions: A survey their applications. Applications of Statistics,(ed. P.R. Krishnaiah). North Holland, Amesterdam, 383-405 (1977).
19. Patil, G. P. and Rao, C. R. Weighted distributions and size-biased of sampling with applications to wildlife populations and human families. Biometrics, **34**: 179-189 (1978).
20. Rosenblatt, M. Remarks on some non-parametric estimates of a density function. Ann. Math. Statist.**27**: 832-837 (1956).
21. Schuster, E. F. Estimation of a probability density and its derivatives. Ann. Math. Statist. **40**: 1187-1196 (1969).
22. Silverman, B. M. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Ann. Statist. **6**: 177-184 (1978).
23. Van Ryzin, J. On strong consistency of density estiamtes. Ann. Math. Statist. **40**: 1765-1772 (1969).
24. Vardi, Y. Nonparametric estimation in renewal processes. Ann. Stat. **10**: 772-785 (1982).
25. Wang, M. C. Nonparametric estimation from cross-sectional survival data. J.Ame. Stat. Asso. **86**: 130-143 (1991).
26. Wicksell, S. D. The corpuscle problem. A mathematical study of a biometric problem. Biometrika, Vol. **17**, No. ½ : 84-99 (1925).
27. Zhang, B. A note on the strong uniform consistency of kernel density estimators under random censorship. Sankhy ā. **60**, Ser. A: 265-273(1998).